Dear Dr. Peleg,

Thank you very much for the thorough assessment of our manuscript and for your invitation to resubmit our manuscript to HESS. Following your and the reviewers' comments, we rewrote the introduction to highlight the aim and novel contribution of the study. The main aim is to 'evaluate the extent to which models calibrated according to standard model calibration metrics such as the widely-used Kling–Gupta efficiency are able to capture flood spatial coherence and triggering mechanisms.' We specify that 'we first evaluate how well the different models capture local flood events following the current paradigm, secondly we expand the evaluation by analyzing how well the models capture spatial flood dependence, and finally we evaluate how the models capture flood triggering mechanisms.' In addition, we added a discussion section that is separate from the results section and expands the previous discussion of the results by discussing potential ways of improving model performance by developing flood-tailored calibration metrics, proposing spatial calibration metrics, identifying model structures representing the most important flood producing mechanisms, and investigating input uncertainty. We think that the major changes made to the introduction, discussion, and conclusions section help to clarify the storyline.
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

Editor's decision

Dear Authors,

I have now received the reports of two referees, one (that did not revise the original text) suggested major revisions, while the other suggested minor revisions (although her/his comments read to me as moderate revisions). Reading the revised manuscript and the comments made by the reviewers, I conclude that additional changes to the text are needed before it can be considered for publication in HESS.

The main issue that I see here, is that the motivation, objectives and hypotheses of the study are not composing a clear storyline. This was already pointed by some of the reviewers in the first round of revision. In this study, a single objective function (KGE) is used for the model calibration, and you demonstrate that 4 different models are failing to represent the observed floods using this single-criteria objective function. In the conclusions, you suggested using multi-criteria objective functions for the calibration of the models if the focus is on representing flood events. This conclusion is not new – many studies in the past used multi-criteria objective functions to calibrate hydrological models to simulate floods, likely with a better match than can be obtained with KGE. Why have you chosen to calibrate the models using an objective function that is known (or can be expected) to fail to simulate flood events to begin with? What multi-criteria objective function/strategy could be used to calibrate hydrological models to better represent flood events (what strategies were used in the past and how they can be improved)? Will a multi-criteria objective function improves the match to flood events, or does some of the models that are presented here will still fail in reproducing flood events due to their internal structure? I am missing answers/discussion to these type of questions.

In my view, **the introduction, discussion and conclusions sections will require considerable text edits** to make the story clearer and more appealing to the readers of HESS, maybe also with minor

changes to the structure of the text. I will be happy to reconsider the revised paper after major revisions.

Reply: *Thank you very much for your clear opinion on how the manuscript can/should be improved. We rewrote the introduction to highlight the aim and novelty of the paper. The main aim is to 'evaluate the extent to which models calibrated according to standard model calibration metrics such as the widely-used Kling–Gupta efficiency are able to capture flood spatial coherence and triggering mechanisms' and the novelty is that 'we expand the evaluation by analyzing how well the models capture spatial flood dependence and how the models capture flood triggering mechanisms.' We focused the study on the Kling-Gupta efficiency metric because it is 'widely used in flood simulation studies (e.g. Hirpa et al., 2018; Huang et al., 2018; Thober et al., 2018; Brunner and Sikorska-Senoner, 2019; Harrigan et al., 2020) and has been shown to result in more accurate flood peak representations than the other widely used individual metric E_{NS} [Mizukami et al., 2019]'.*
With this study, another key aim is to raise awareness that notwithstanding the popularity of the KGE calibration metric in recent years, it may not be the best choice if one is interested in floods. This outcome may be appreciated by a small portion of the field, but we do not see evidence that it is widely appreciated, either from the literature or in the authors' own interactions with other researchers. There has, on the whole, been much aspirational discussion of more widespread use of hydrologic signatures, but there is also a continuing practice of defaulting to the generic KGE or NSE for many studies. These notably include some of the studies by the authors themselves, which in part produced these parameter sets. We added a new discussion section that significantly expands the discussion on potential multi-criteria objective functions for flood events, spatial model calibration, and the role of model structure in model performance. We think that the rewritten introduction, discussion, and conclusions sections convey a clear and useful story to the HESS readership.

Reviewer 1

General comments

This study compares the efficiency of three lumped and one distributed models to simulated flood magnitude, timing and spatial coherence. The objective function used for calibration is Kling-Gupta efficiency (KGE). The results show that models tend to underestimate flood magnitude and not always simulate well flood timing. The authors conclude that using KGE for calibration has limited reliability for flood hazard assessment.
In general, the topic fits scope of the journal and will be of interest for the readers. However, the manuscript in its current form (after the revision) will still benefit from a more thorough revision. The main critical points (in my opinion) are:

1) The formulation and justification of the novel scientific contribution is still not clear. The review of previous studies in the Introduction indicates that "…to achieve further improvements in flood peak simulations, a broader range of application-specific evaluation metrics is typically required." (l.29-30, l.23-24). I agree with such formulation of current research gaps, but it is not in line with the objective function tested in the manuscript. If one would be interested in flood magnitude, timing and spatial connectivity, why one should use KGE for calibration? How does it account for such specific evaluation metrics, i.e. flood seasonality or spatial coherence?

Reply: *Thank you for highlighting the need to better work out the novelty of our study and to justify the use of E_{KG} for model calibration. The main aim is to 'evaluate the extent to which models calibrated according to standard model calibration metrics such as the widely-used Kling-Gupta efficiency are able to capture flood spatial coherence and triggering mechanisms' while the novelty is that 'we expand the evaluation by analyzing how well the models capture spatial flood dependence and how the models capture flood triggering mechanisms.' We focused the study on the Kling-Gupta efficiency metric because it is 'widely used in flood simulation studies (e.g. Hirpa et al., 2018; Huang*

et al., 2018; Thober et al., 2018; Brunner and Sikorska-Senoner, 2019; Harrigan et al., 2020) and has been shown to result in more accurate flood peak representations than the other widely used metric E_{NS} [Mizukami et al., 2019]'. We chose E_{KG} for calibration to highlight that the widely used metric might not lead to accurate flood simulations as often assumed. The point is exactly that E_{KG} is often used as a metric in flood simulation studies despite the fact that it may lead to suboptimal model performance with respect to floods. With this study, we want to raise awareness for exactly this issue. By raising awareness of this matter, we hope to inspire other researchers to propose alternative calibration metrics. In the newly created discussion section, we present ideas on how flood simulations could be improved by developing alternative calibration metrics, developing spatial calibration metrics, identifying suitable model structures, and improving precipitation input data.
Modification: p1. l.5-8, p.2 l.27-34 and l.49-54

2) The title is misleading. The main message of the paper, in its current form, is about the value of KGE for calibration of hydrologic models (if flood impact assessment is the main purpose). There is no assessment how the models describe and simulate different flood generation processes and which factors control their performance. So based on presented results it is difficult to interpret to what extent and how are the selected models suitable for flood impact assessment. The results are more about the accuracy of selected way (i.e. using lumped models, KGE for calibration, etc.).
Reply: *Thank you for pointing out the need for revising the current title to better reflect the main message of the study. The revised version goes beyond discussing the value of E_{KG} as a calibration metric by also discussing the role of model structure, which is enabled by the comparison of the performance of four different models. We chose the following new title: 'Flood spatial coherence, triggers and performance in hydrological simulations: large-sample evaluation of four streamflow-calibrated models', which highlights that the paper is about model evaluation for floods and reflects the focus on spatial flood characteristics and the representation of flood drivers.*
Modification: title

3) The significance of the results is not clear. I'm not sure if for practical applications, a lumped model will be used or should be recommended. Perhaps a consistent assessment/evaluation of the difference between lumped and distributed type of models will be interesting (e.g. for HBV and mHM).
Reply: *We agree that a more in-depth discussion of the results was needed in order to highlight their significance. We therefore separated the Results section from a newly created Discussion section. In this new section, we discuss the findings and propose potential ways of improving flood simulations by moving away from standard calibration metrics such as E_{KG}, by identifying suitable model structures, and by improving the quality of input precipitation. We agree that an explicit assessment of the role of model type (lumped vs. distributed) would be interesting and we think that such an assessment is out of scope of this paper.*
Modification: p.12 l.232-p.17 l.329

4) The design of the experiment reads more as a collection of available analyses and not results from initially clearly defined research question/hypothesis. I agree with previous reviews that using different time periods for calibration and using different model input datasets can have some impact on the results and the interpretation of results (including individual catchments) will be more consistent if the same data and time periods will be used. The authors claim that both datasets describe the observed climate, but are they identical also for individual extreme events?
Reply: *We reworked the introduction to highlight the main aim of the study, i.e. 'This study evaluates the extent to which models calibrated according to standard model calibration metrics such as the widely-used Kling--Gupta efficiency are able to capture flood spatial coherence and triggering mechanisms', and stress that 'we expand the evaluation by analyzing how well the models capture spatial flood dependence and how the models capture flood triggering mechanisms.' To achieve this goal, we use simulations generated in previous studies, which serve as a proxy for simulations that*

would be typically used in research studies on flooding because they used best possible calibration settings given past computer and resources availability. We agree that ideally the same precipitation input would have been used for all models, which was unfortunately not possible because the simulations our study is based on were derived by different authors. The two datasets, however, were derived from observed precipitation and temperature, and have been shown to result in similar mean daily precipitation fields [*Newman et al.*, 2015]*. We therefore consider them similar enough to allow for a direct comparison of the model outputs resulting from the different input datasets.*
Modification: p.2 l.49-54

5) The methodology is not rigorously described. It will be very difficult (if even possible) to reproduce/repeat the presented analysis (based on given information). Numerous information is missing, e.g., how the initial values were set, what were the ranges of calibrated model parameters and parameters of automatic calibration algorithm. It will be interesting to present, e.g. in appendix, the final model parameters and efficiencies for individual catchments. This will allow to assess the interpretation made.
Reply: *Our study seeks to extract information through the pooling of different modeling results. For this, we used streamflow simulations derived in previous studies by Melsen et al. (2018) and Mizukami et al. (2019) as described in the Methods section. Thus, we did not have complete control over the design of the experiment. As specified in the data availability section, model simulations can be obtained by the main authors of these previous studies. Details on the model calibration procedures used in these studies can be found in the references. Melsen et al. (2018) provide information on the parameter boundaries used in the Sobol-based Latin hypercube sampling in Tables C1-C3. Mizukami et al. (2019) did not provide specific parameter ranges in the original paper published. We think that providing model parameters for 671 catchments is infeasible even in the appendix as this would produce a large amount of additional pages, which in our opinion hardly anyone would look at.*

6) I think that comparing lumped with distributed models can bring some more interesting results than are presented in its current form. What is the impact of lumping on the results? Are the differences in model efficiency related to the size of the basin? I would expect that using lumped models in larger catchments cannot describe well floods from convective rainfalls.
Reply: *Thank you for suggesting this additional analysis. Figure 1 shown in this review shows flood model errors at individual sites for five different catchment size classes. Model performance does not seem to depend on catchment size and we can not identify significant differences in outcomes between distributed models (mHM) and lumped models (other three models). The fact that we can draw similar conclusions from both lumped and distributed models strengthens the argument being made. The lumped/distributed impact questions would be a good topic for a follow on study, though there is some literature on that topic already (e.g the Distributed Model Intercomparison Project DMIP study: https://www.weather.gov/owp/oh_hrl_distmodel_dmip_draft).*

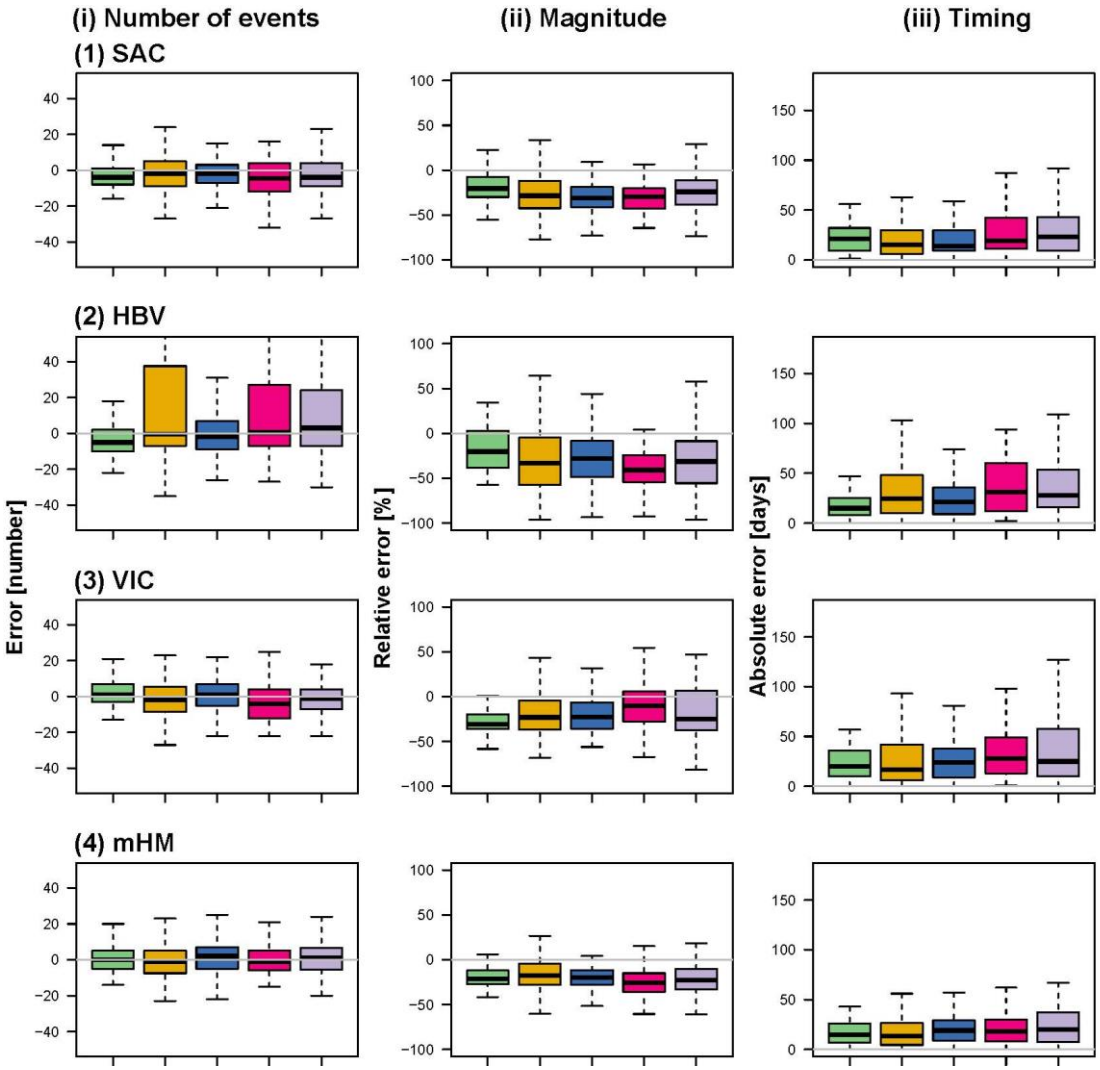

Figure 1: Model errors per catchment size class computed over the period 1981-2008: <100 km^2 (green, 93 catchments), >=100 and <250 km^2 (yellow, 115), >=250 and <500 km^2 (blue, 112), >=500 and <1000 km^2 (pink, 93), and >1000 km^2 (violet, 75). Errors are shown for (i) number of events (error in number of events), (ii) magnitude (mean relative error in %), 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 catchment size class.

7) As a reader, I would be likely more interested in seeing where (in which catchments and why?) the models work well, rather than to conclude that in general they underestimate magnitude or do not represent well the timing or spatial patterns. So presenting some deeper analysis of the factors controlling the performance will be helpful and interesting.

Reply: *We agree that providing model evaluations for different types of catchments is interesting. We therefore perform model evaluations for 5 types of streamflow regimes (see Figure 1) as shown in Figures 2, 3, 4, and 7. This regime-specific analysis allows us to conclude that 'model performance is generally worst in catchments with intermittent regimes while it is highest for catchments with a strong seasonality such as a melt and New Year's regime.' We highlight the regime-specific analysis in the Methods section by saying: 'To provide insights with respect to where model performance is better/worse, we provide model evaluation results for five different streamflow regime types, which have been shown to be distinct in their flood behavior: 1) Intermittent, 2) weak winter, 3) strong winter, 4) New Year's, and 5) melt. Catchments with intermittent regimes experience floods mainly in*

spring and summer, those with weak winter regimes in winter and spring, those with strong winter regimes in winter, those with a New Year's regime around New Year, and those with a melt-dominated regime in spring because of snowmelt.'
Modification: p.4 l.95-100

Specific comments

1) Abstract, l.13: "…models calibrated on integrated metrics such as …have limited reliability…". This is a general conclusion which is not supported well with the presented results. I would suggest to remove "such as". I think if one combines flood magnitude, seasonality and spatial coherence into an integrated metrics (objective function) for calibration, the results can be different.
Reply: *Thank you for the rephrasing suggestion, which we adopted.*
Modification: p.1 l.16

2) Data. I like the assessment based on large dataset and subsequent split/grouping of results into some relevant groups of catchments. It is however not clear how are flood generation processes (e.g. flood types) linked with selected groups of regimes? If the objective is about the suitability of models to represent floods (magnitude, seasonality, …) it will be interesting to see results for different flood generation mechanisms, i.e. how or if the models differ in simulating snowmelt floods, or floods from convective rains, etc.
Reply: *Thank you for pointing out to better introduce the regime-specific analysis. We specify in the Methods section that 'To provide insights with respect to where model performance is better/worse, we provide model evaluation results for five different streamflow regime types, which have been shown to be distinct in their flood behavior: 1) Intermittent, 2) weak winter, 3) strong winter, 4) New Year's, and 5) melt (Figure 1; Brunner et al.*, 2020*). Catchments with intermittent regimes experience floods mainly in spring and summer, those with weak winter regimes in winter and spring, those with strong winter regimes in winter, those with a New Year's regime around New Year, and those with a melt-dominated regime in spring related to snowmelt.' We agree that further distinguishing between different flood generation types would be very interesting as well. However, we think that such an assessment would be a separate (classification/clustering) study in itself.*
Modification: p.4, l.95-100

3) Forcing. Which version of Daymet is used? Why not to use only one dataset for all the models?
Reply: *We based this study on previously published work, which used best possible calibration settings given past computer and resources availability, and scope. These studies used slightly different model inputs as described in the Methods section. The two datasets, however, were derived from observed precipitation and temperature, and have been shown to result in similar mean daily precipitation fields* [*Newman et al.*, 2015]*. Melsen et al.* (2018) *used Daymet version 2.1. for their simulations with SAC, HBV, and VIC. Recognizing some inconsistencies in the different sources of data, we nonetheless felt that extending our analysis to cover the multiple models would make our findings more robust.*

4) The term "event": By using term flood event, do you mean day of the flood peak? The same for precipitation. Is the event precipitation representing mean daily precipitation for the day of the peak? Some flood events (e.g. from snowmelt) can last several days. How sensitive/representative are the characteristics extracted only for the day of the peak?
Reply: *Thank you for highlighting the need for clarifying the meaning of the term 'event' and for specifying how corresponding rainfall, snowmelt, and soil moisture were identified. We specify that by a peak-over-threshold flood event, we mean peak discharge, and that precipitation, snowmelt, and soil moisture were identified for the day of peak discharge. To identify how sensitive the results shown in Figure 5 are to the aggregation level (i.e. 1 day), we performed the same analysis also with 3-day precipitation and snowmelt sums. The results look almost identical to the ones presented for*

the 1-day aggregation level.
Modification: p.6 l.119, p.7 l.160

5) Beta model parameter (l.170). It will be interesting to present model parameters for individual catchments, because otherwise the interpretation made reads more as speculation (it is not justified by presented results).
Reply: *Thank you for indicating the need to distinguish between results and their discussion. To do so, we created a new Discussion section. We moved the statement about the role of the beta parameter to this new discussion section to highlight that we use it to interpret the results presented in Figures 3 and 4.*
Modification: New discussion section (p.12-17)

6) L.218-219. In my opinion HBV model can describe the surface runoff. Conceptual it is represented by the outflow from the upper reservoir (describing by k0 model parameter).
Reply: *We think that this statement is correct because [Bergström, 1976] wrote: 'All versions of the HBV-model are lacking components for direct surface runoff, as the water is controlled by the conditions in the soil moisture zone before any runoff can be generated.'*

Reviewer 2

The authors responded to most of my concerns, but some issues are left.
Reply: *Thank you very much for taking the time to write this second review.*

[1] The title is still problematic. As far as I can tell, there is no "flood impact assessment" performed in this study. Why is it in the title? The study assess flood flows, so why is this not the title of the paper? Flood impact assessment would require a direct connection to the actual implication of flooding, such as flood inundation, damage to houses etc. These aspects are not part of the study, so why is the title focusing on this issue?
And, if the focus is on assessing the value of KGE as calibration metric, then why is this not in the title? The title "Evaluating the suitability of hydrological models for flood impact assessments", is still much broader than what this very focused study actually does.
Reply: *Thank for pointing out the need to further improve the title. We agree that we are not performing a flood impact assessment and that talking about flood simulations instead would be more appropriate. We propose the following revised title: 'Flood spatial coherence, triggers and performance in hydrological simulations: large-sample evaluation of four streamflow-calibrated models', which stresses that this study is about model evaluation and that the focus is on both local and spatial flood characteristics.*
Modification: Title

[2] (line 140) The use of split sample schemes should include a reference back to Klemes (1986, HSJ, https://www.tandfonline.com/doi/abs/10.1080/02626668609491024) who introduced the idea.
Reply: *Thank you for indicating the need to cite the original reference to the split sample testing idea. We added the reference to the text.*
Modification: p.2 l.48, p.7 l.154

[4] Section 3.1: I asked previously why HBV results are so poor and I am still confused by it. It would be useful for the discussion section of this paper to more closely compare the results obtained here to previous studies across the USA. For example, Kollat et al. (2012, WRR, https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2011WR011534) calibrated the HBV model across all MOPEX catchments and found Nash Sutcliffe Efficiency values much higher than what would be expected based on the results of the current study (see their Figure 9A). Why the discrepancy? Kollat et al. (2012) performed extensive MO-calibration whereas the current study used

a LHS sampling strategy. So, is part of the result of the current study is due the chosen calibration approach?

Reply: *Thank you for indicating this reference. We looked the E_{NS} values presented in Figure 9A of Kollat et al. (2012). This figure shows that roughly 50% of the catchments show a E_{NS} value >0.75. If we determine the percentage of catchments in our dataset that has E_{KG} (we did not use E_{NS} to identify best parameter sets) values >0.75, we get 47%. Similarly, roughly 90% of the catchments in the Kollet et al. (2012) study show E_{NS} values above 0.5, the same percentage of catchment that also exceeds E_{KG} values of 0.5 in our study. We therefore argue that the model performance of the HBV model used in our study is as good as the performance of the model calibrated in the Kollat et al. (2012) study.*

[5] Other studies have disaggregated KGE to understand what controls the bias in the KGE terms. E.g. Gudmundsson et al. (2012, WRR, https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2011WR010911) found that one significant control on water balance error seemed to be precipitation data error – model predictions in catchments with significant elevation difference (where at least across Europe, precipitation measurements are expect to be less good) were performing poorer. Did the authors find similar patterns? Would elevation difference be a good way to see whether rainfall is indeed a likely problem for the study catchments in this paper?

Reply: *As suggested, we checked whether model performance in terms of E_{KG} is related to elevation (Figure 2 shown in this response to the reviewers). High-elevation catchments have generally better performance than low elevation catchments, but present a confounding factor in that higher catchments more often experience snow, which has generally a positive influence on model calibration. This finding suggests that precipitation errors, which are typically higher in high-elevation catchments due to less dense measurement networks, are not the main determinant of model performance.*

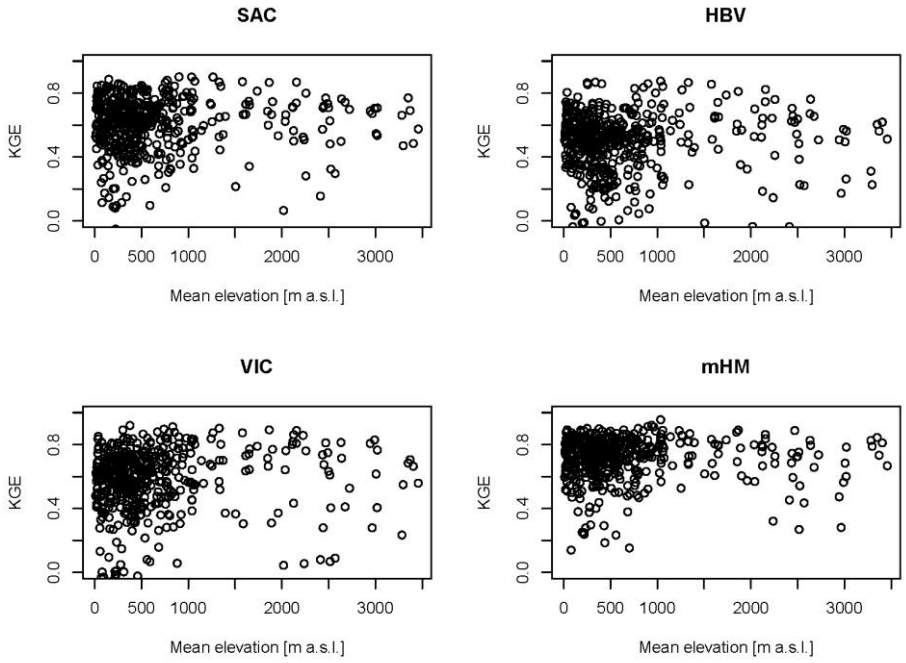

Figure 2: E_{KG} vs. elevation for the four models tested: SAC, HBV, VIC, and mHM.

[3] I am still confused by the authors conclusion that "Our model comparison shows that all flood characteristics are not equally well represented by models calibrated with the widely used Kling–Gupta efficiency metric." – OK, but very likely this is true for any metric given the extensive experience with multi-objective model calibration in hydrology, where a regular finding is that any single metric produces a focused result. So, what multi-objective strategy do we need to improve this

problem? And what relevant trade-offs exists (e.g. Kollat et al., 2012)? The authors suggest that multi-objective calibration is the way forward, in which case a better review of this very rich multi-objective literature in hydrology would be nice (given that this topic has been explored for over 20 years). Currently there are only a couple recent references, which do not do the topic justice – even if narrowed down to those studies focusing on calibration for flood prediction.
Reply: *Thank you for highlighting the need to expand the discussion on multi-objective calibration strategies. We significantly expanded the discussion of multi-objective calibration in the context of flood modeling and also added references to the more classical literature on multi-objective calibration not necessarily targeted at floods.*
Modification: p.16 l.282-295

The next sentence suggests a much wider conclusion: "The number of floods, flood magnitude, and timing are not always well captured by hydrological models in many catchments." It would be good if the authors were to formulate their conclusions more carefully. Given that the authors have a very narrow focus in this study (which is fine) – to show that calibrating to KGE does not lead to a good reproduction of all flood characteristics – it would be good to formulate their conclusions with a similar focus to avoid that others misuse their conclusions.
Reply: *Thank you for pointing out the need to more concisely phrase the conclusions. We changed this sentence to: 'Our model comparison shows that flood characteristics are not always well captured in hydrological models developed for research studies – even when the models have been calibrated with a calibration metric perceived suitable for flood modeling, the Kling–Gupta efficiency metric (EKG).'*
Modification: p.17 l.331-333

Flood spatial coherence, triggers and performance in hydrological simulations: large-sample evaluation of four streamflow-calibrated models.

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. A reliable hydrologic model ideally reproduces both local flood characteristics and spatial aspects of flooding under current and future climate conditions. However, uncertainties in simulated floods can be considerable and yield unreliable hazard and climate change impact assessments. This study evaluates the extent to which models calibrated according to standard model calibration metrics such as the widely-used Kling–Gupta efficiency are able to capture flood spatial coherence and triggering mechanisms. To highlight challenges related to flood simulations, 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 also in flood-related studies. Specifically, 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. From a large sample of catchments and with multiple models, we conclude that calibration on the integrated Kling–Gupta metric alone is likely to yield models that have limited reliability in flood hazard assessments, undermining their utility for regional and future change assessments. We underscore that such assessments can be improved by developing flood-focused, multi-objective and spatial calibration metrics, by improving flood generating process representation through model structure comparisons, and by considering uncertainty in precipitation input.

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 model ideally reproduces different aspects of flooding, including local characteristics such as event magnitude and timing. To obtain such satisfactory flood simulations, hydrological models are often calibrated using one or several objective functions. One widely-used metric that is often used in flood studies (e.g. Hundecha and Merz, 2012; Köplin et al., 2014; Vormoor et al., 2015; Wobus et al., 2017) is the Nash–Sutcliffe efficiency (E_{NS}; Nash and Sutcliffe 1970) because it is considered integrative compared to others and focuses attention on high flows. However, E_{NS} is formulated so that its optimal value systematically underestimates flow variability (Gupta et al., 2009), undermining the ability of a model to reproduce peak flow values. A related metric, the Kling–Gupta efficiency (E_{KG}; Gupta et al. 2009), is free from this constraint and may improve simulations of peak flows, especially if the variability related component of the score is emphasized in calibration (Mizukami et al., 2019). This metric has been frequently used in recent flood modeling studies (e.g. Harrigan et al., 2020; Hirpa et al., 2018; Huang et al., 2018; Thober et al., 2018; Brunner and Sikorska, 2018) and seems to be widely accepted as a suitable choice for flood studies. This may arise from the general practice of developing models for a range of objectives. However, recent studies have shown that capturing flood magnitude and timing is challenging when such standard calibration metrics are used for parameter estimation (Lane et al., 2019; Brunner and Sikorska, 2018; Mizukami et al., 2019).

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.

Local and spatial flood characteristics should be reliably simulated not only under current but also under future climate conditions. However, models calibrated for current conditions may not be transferable in time (Thirel et al., 2015) partly because of a sub-optimal representation of flood producing mechanisms. To overcome this transferability problem, the differential split-sample test has been proposed, where the model is calibrated and validated on two periods with differing climate conditions (Klemes, 1986; Seibert, 2003).

In this study, we evaluate the extent to which model calibrated according to the widely-used model calibration metric E_{KG} are able to capture flood spatial coherence and flood triggering mechanisms. To this end, we first evaluate how well different hydrological models capture local flood events following the current paradigm, secondly we expand the evaluation by analyz-

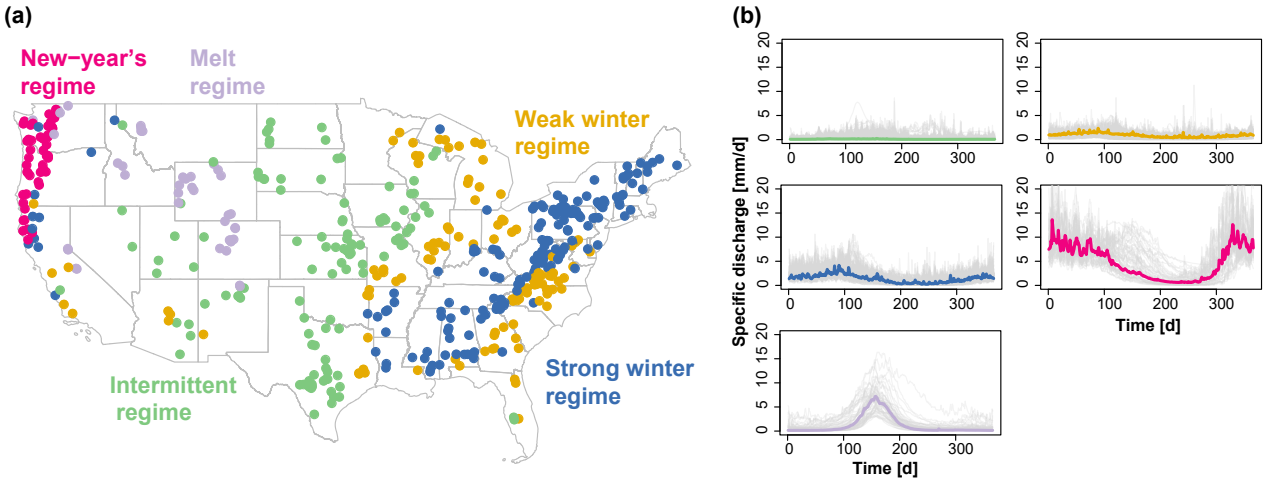

Figure 1. a) Map of the 488 catchments in the conterminous United States belonging to the five regime classes indicated by their gauge location: 1) Intermittent, 2) weak winter, 3) strong winter, 4) New Year's, and 5) melt. b) Median regime per regime class (colored lines) and variability of regimes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

ing how well the models capture spatial flood dependence, and finally we evaluate how the models capture flood triggering mechanisms. With this thorough evaluation, 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.

For documenting modeling challenges related to floods, 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, 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., 2015b). 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 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 (1 km resolution; Thornton et al., 2012) and mHM with the forcing by Maurer et al. (2002) (12 km 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.

To provide insights with respect to where model performance is better/worse, we provide model evaluation results for five different streamflow regime types, which have been shown to be distinct in their flood behavior: 1) Intermittent, 2) weak winter, 3) strong winter, 4) New Year's, and 5) melt (Figure 1; Brunner et al., 2020b). Catchments with intermittent regimes experience floods mainly in spring and summer, those with weak winter regimes in winter and spring, those with strong winter regimes

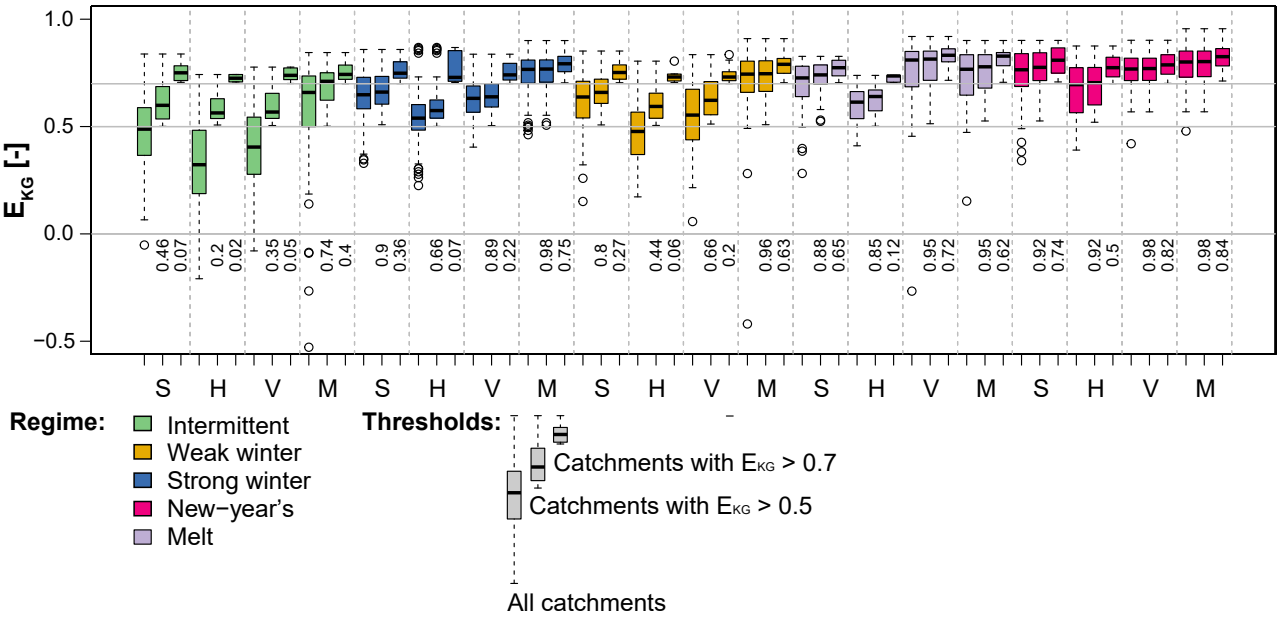

Figure 2. Model performance in terms of E_{KG} over the period 1981–2008 for the four models SAC (S), HBV (H), VIC (V), and mHM (M) per hydrological regime: intermittent (114 catchments), weak winter (108), strong winter (176), New Year's (50), and melt (40). For each model and regime, three boxplots 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.

in winter, those with a New Year's regime around New Year, and those with a melt-dominated regime in spring because of snowmelt.

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.

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 for the five streamflow regimes introduced above. Such a comparison enables

identification of flood characteristics whose model representation could potentially be improved. To better understand potential model deficiencies, we look at how models capture flood triggering mechanisms and how they simulate floods under climate conditions different from the current ones.

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 (i.e. peak discharges) 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 25^{th} 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 of 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: $n_\text{s} - n_\text{o}$, where s represents simulations and o observations), magnitude in terms of mean peak discharge x (relative error: $(x_\text{s} - x_\text{o})/x_\text{o}$), 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 (Klemes, 1986; 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 on the day of peak discharge. 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). Event magnitude in terms of peak discharge is generally underestimated for all regime types independent of the model and also 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 a few months (Brunner et al., 2020a).

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. Connectedness overestimation by HBV is most pronounced for catchments with an intermittent regime. Otherwise, connectedness over-/underestimation seems to be independent of the regime.

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.

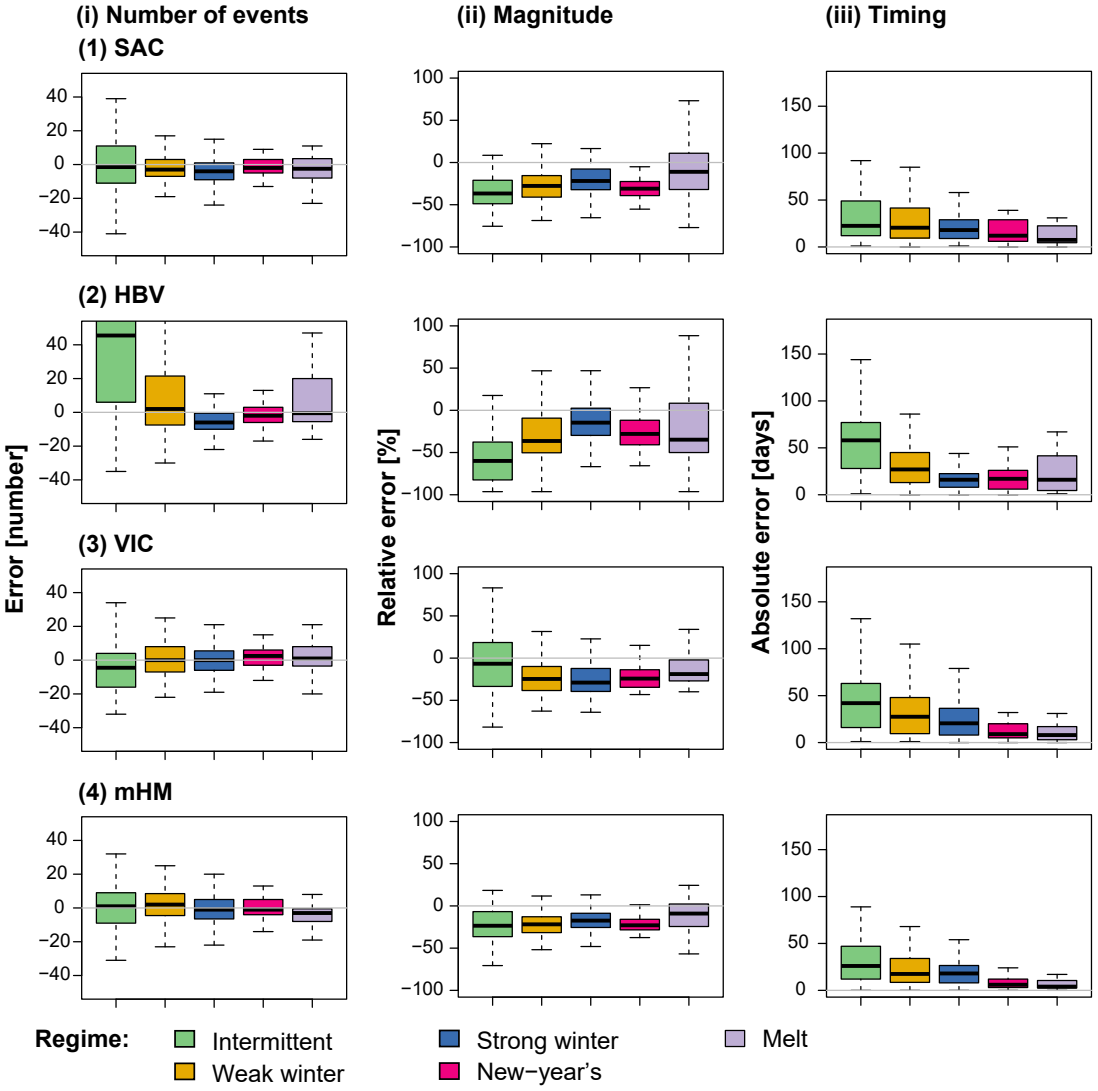

Figure 3. Model errors per regime type computed over the period 1981–2008: intermittent (114 catchments), weak winter (108), strong winter (176), New Year's (50), and melt (40) (Figure 1). Errors are shown for (i) number of events (error in number of events), (ii) magnitude (mean relative error), and (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.

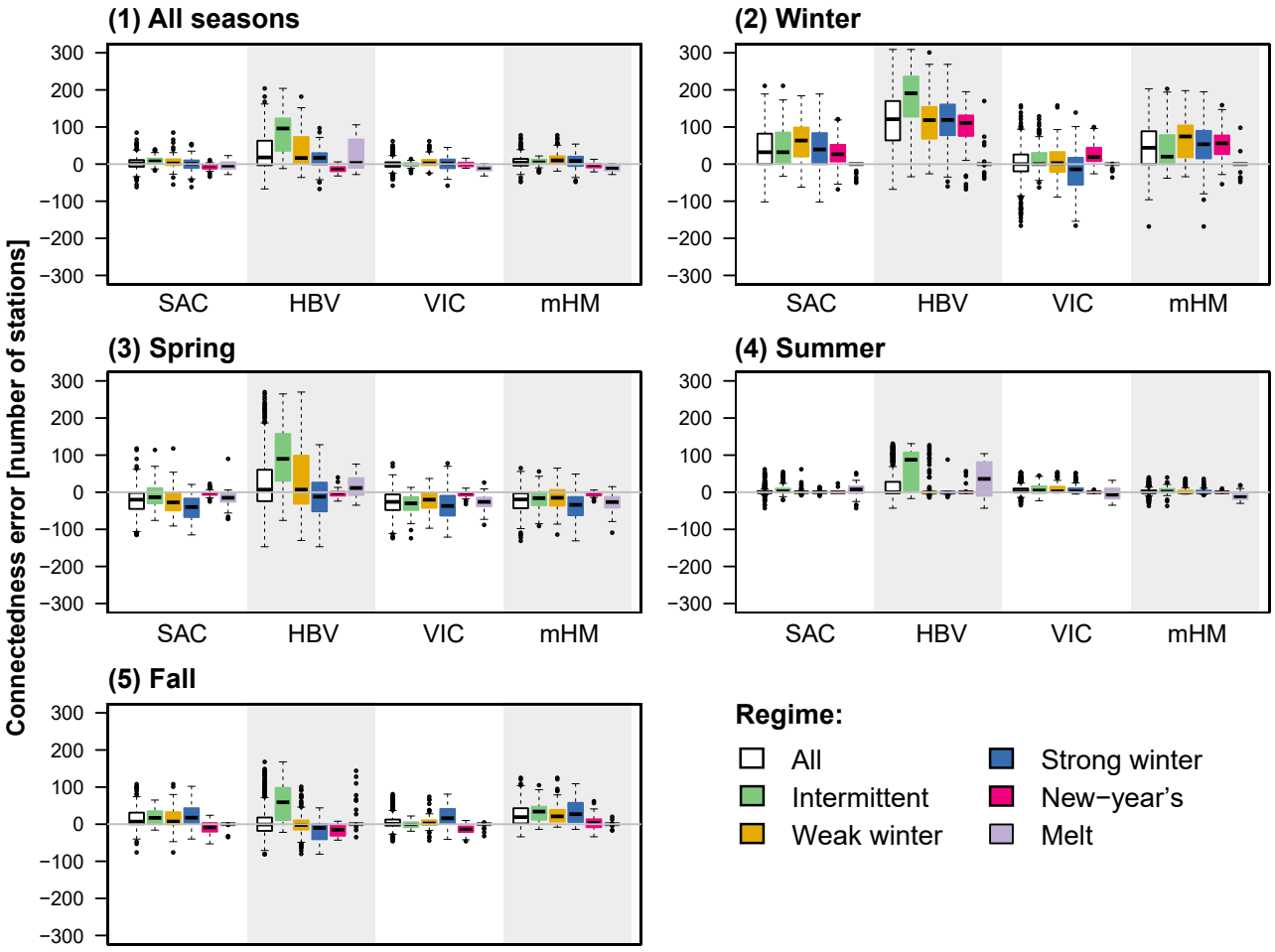

Figure 4. Overall (1) and seasonal (2–5) errors in flood connectedness (simulated minus observed connectedness), i.e. number of catchments a catchment is sharing at least 1% of the total number of flood events with, for the four models SAC, HBV, VIC, and mHM over all regimes and per regime: intermittent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

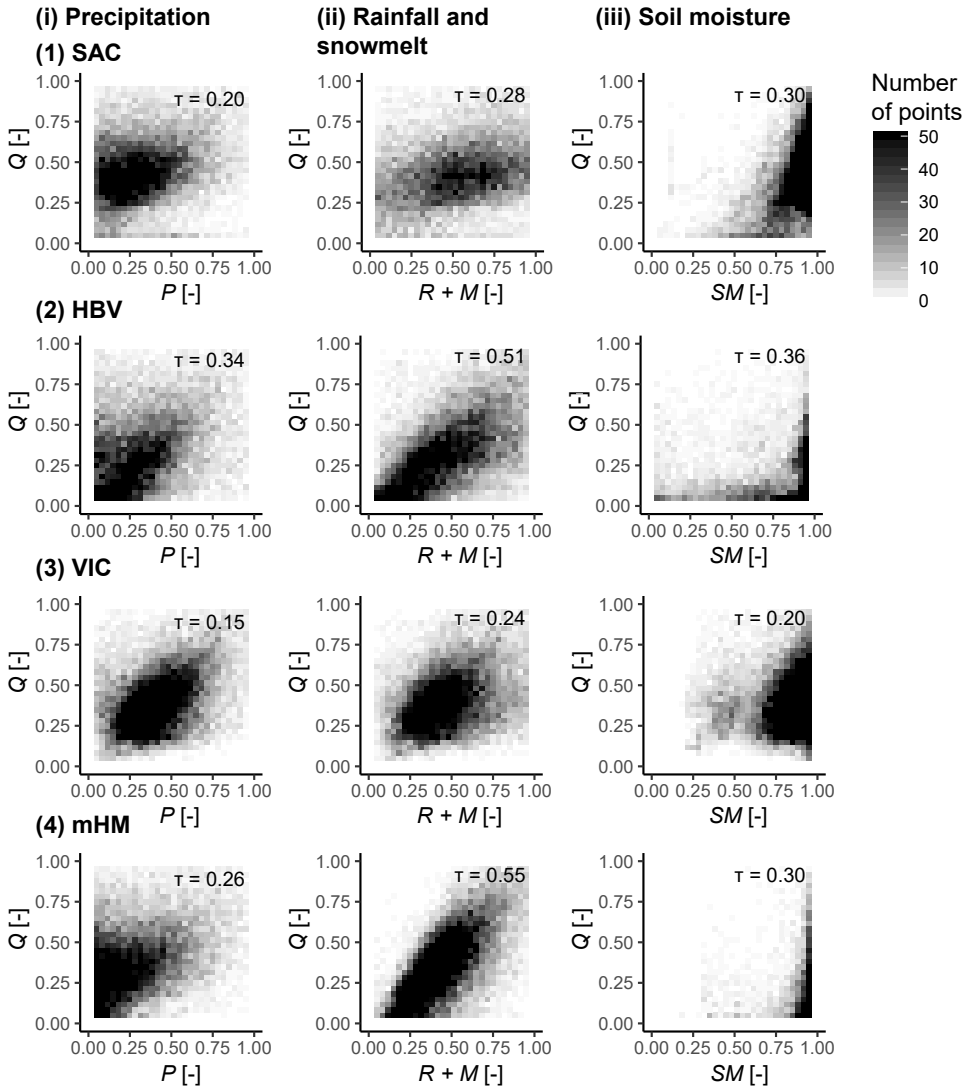

Figure 5. Simulated relationships between normalized flood discharge (Q) and normalized precipitation (i, P), rainfall and snowmelt (ii, $R+M$), and soil moisture (iii, SM, upper two soil layers for mHM) over all catchments represented by a binned scatter plot for the four hydrologic models (1) SAC, (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.

However, peak discharge and rainfall plus snowmelt show a strong linear relationship, i.e. the higher the combined rainfall and 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.

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 model's 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 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.

4 Discussion

4.1 Model performance in simulating floods

The results presented in this study demonstrate that simulating floods using hydrological models is challenging both at a local and spatial scale. At the local scale, flood timing and magnitude may not be perfectly captured which can translate into a sub-optimal representation of spatial dependencies because space and time are closely related. The challenges related to flood

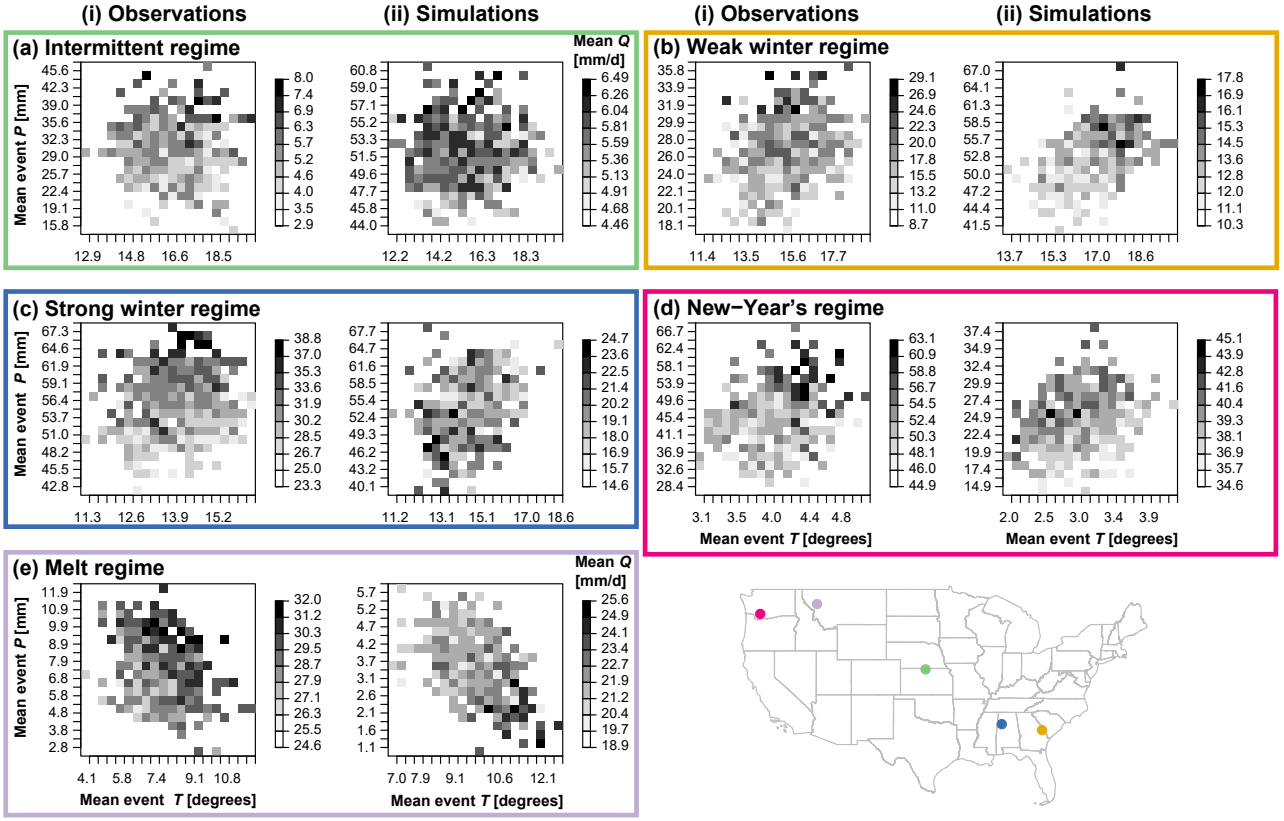

Figure 6. Climate sensitivity analysis for the VIC model: Dependence of mean POT magnitude (Q) on mean flood event precipitation (1-day; P) and mean flood temperature (T) for five example catchments, those with the best E_{KG} per regime type: intermittent regime (green; USGS ID 09210500 Fontanelle Creek 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.

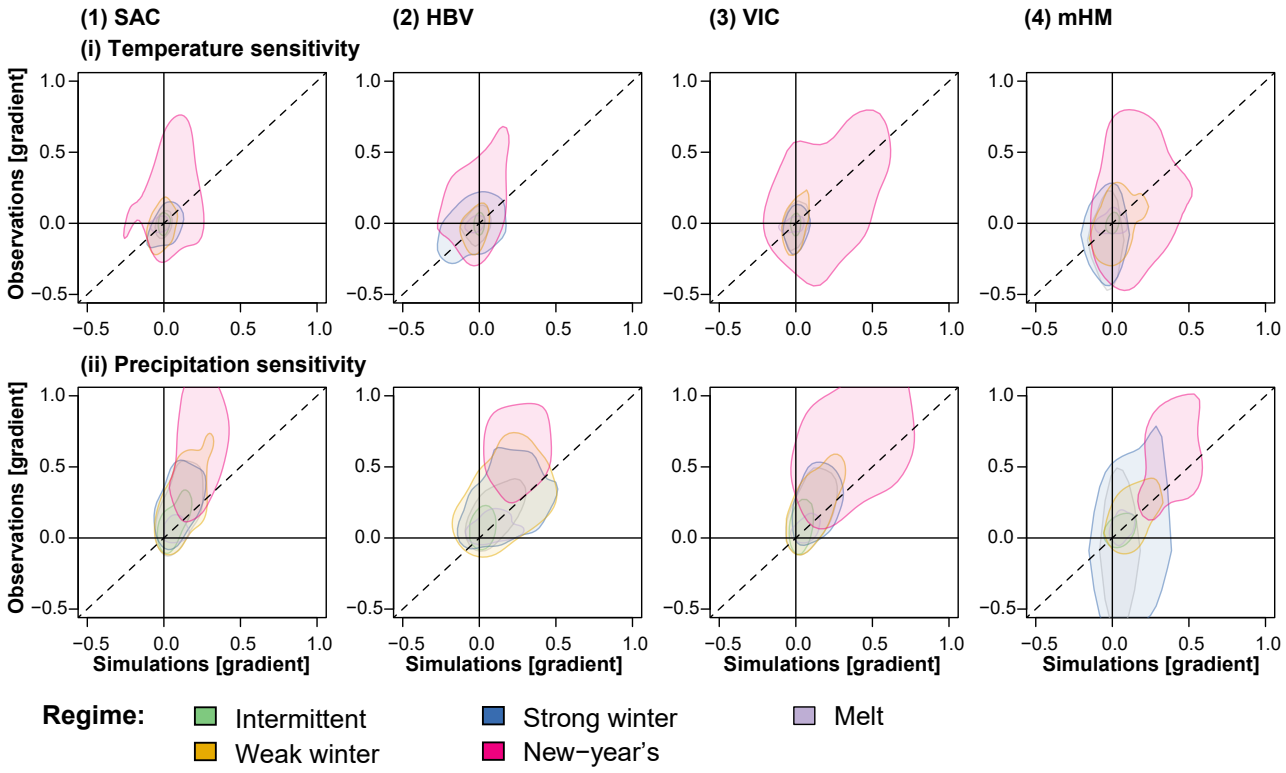

Figure 7. Observed vs. simulated (i) horizontal (temperature) and (ii) vertical (precipitation) climate sensitivities for floods represented by two-dimensional kernel density estimates for the four models (1) SAC, (2) HBV, (3) VIC, and (4) mHM for the five regime types: intermittent (114 catchments), 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.

simulations become especially pronounced under climate conditions different from the current ones because additional sources of uncertainty are added to the modeling chain.

Even though the models have been calibrated for the local situation, substantial differences in magnitude and timing were found between observations and simulations. Locally, simulated floods showed smaller magnitudes and had different timing than observed ones while the number of floods was reproduced relatively well except by the HBV model for catchments with intermittent regimes. The flood magnitude underestimation found for all four models tested 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 have smoothed the simulated discharge response.

Under the current calibration paradigm, where models are calibrated to local discharge conditions using E_{KG} as objective function, flood conectedness is not accounted for. As a result, flood contectedness is not well captured by the models as illustrated by the finding that flood connectedness is over- or underestimated depending on the season. 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. The finding that there is room to improve 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).

There are slight variations in performance among models. These variations may result from differences in the representation of flood producing mechanisms as indicated by distinct behaviors in how the models translate precipitation into runoff. 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, the latter which is based on the HBV model structure (Kumar et al., 2013), does 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. These differences in process representation suggests that a 'most suitable model' could be identified for a specific application at hand. If one is e.g. interested in simulating floods in catchments with intermittent regimes, the HBV model does not seem to be an ideal choice because it there simulates too many floods with a too small magnitude. The overestimation of the number of events in catchments with intermittent regimes 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).

Our climate sensitivity analysis shows that the simulation of floods becomes even more challenging under climate conditions different from the current ones as the hydrological models employed 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), insufficient model calibration (Fowler et al., 2016), or 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).

4.2 Potential ways to improve model performance

The results of our model comparison highlight that there is room for improvement regarding the representation of local flood events, spatial flood dependence, and flood producing mechanisms. We here discuss four potential ways for improving model performance: developing flood-tailored calibration metrics, considering spatial aspects in model calibration, improving representation of flood processes, and representing input uncertainties.

A first possibility to improve model performance is to develop calibration metrics tailored to flooding instead of relying on E_{KG}. Our results show that E_{KG} can lead to simulation performance deficits for phenomena of interest, including an underestimation of peak flow, a misrepresentation of timing, and over- or underestimation of seasonal spatial flood connectedness. 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 representation of streamflow variability could potentially be improved by giving more weight to the variability component of an integrative metric such as the E_{KG} (Pool et al., 2017); whereas the representation of flood magnitude and timing may be improved by giving more weight to the bias and correlation components of the E_{KG}. Alternatively, these characteristics could be optimized explicitly by minimizing error in key hydrograph signatures related to site-specific flood phenomena. Such flood-focused optimization may similarly to E_{KG} rely on multiple objectives in a scalar function (Gupta et al., 1998; Efstratiadis and Koutsoyiannis, 2010) such as volume error, root-mean-squared error, and peak flow error (Moussa and Chahinian, 2009); E_{NS} and relative peak deviation (Krauße et al., 2012); E_{KG}, peak efficiency, and logarithmic efficiency (Sikorska et al., 2018); or E_{KG}, peak efficiency, and mean absolute relative error (Sikorska-Senoner et al., 2020). In addition, model performance can potentially be improved by using multiple metrics describing important catchment processes (Madsen, 2003; Dembélé et al., 2020), i.e. flood generating mechanisms such as soil moisture and snowmelt.

A second way to improve model performance is to focus on the spatial representation of extremes, which may be improved by considering spatially distributed features of model response or spatial correlation within a spatial calibration framework. Such a framework could build upon existing spatial verification metrics such as the spatial prediction comparison test used e.g. to validate precipitation forecasts (SPCT; Gilleland, 2013), Empirical Orthogonal Functions (EOFs), or Kappa statistics (Koch et al., 2015). For the calibration and evaluation of spatially-distributed hydrological models, Koch et al. (2018) recently proposed the SPAtial Efficiency (SPAEF) metric which reflects three equally weighted components: correlation, coefficient of variation and histogram overlap. To improve the spatial dependence of floods across different sites, such spatial calibration frameworks would need to include spatial verification metrics focusing at extremes, which could e.g. be achieved by looking at deviations of simulated from observed F-madograms, which measure extremal dependence (Cooley et al., 2012). Please note, however, that even the use of spatial verification metrics may not overcome the lack of spatial heterogeneity in precipitation or soil moisture data.

A third way of improving model performance is to test whether a model is fit-for-purpose and to identify model structures which accurately represent relevant flood producing mechanisms. The importance of model structure choice has been highlighted in previous studies both for low- and high flow events (Melsen and Guse, 2019; Kempen et al., 2020; Knoben et al.,

2020) and should depend on the spatial complexity of the phenomenon studied (Hrachowitz and Clark, 2017). However, model structure choice for a specific application is not straightforward and automatic model structure identification frameworks have only been introduced very recently (Spieler et al., 2020). To improve the representation of flood processes, such frameworks would ideally explicitly consider local and spatial flood characteristics and the representation of different flood generation processes such as rain-on-snow events or flash floods. The representation of rain-on-snow floods for example requires an accurate representation of the energy balance in order to represent factors affecting snowmelt processes such as net radiation and turbulent heat fluxes (Pomeroy et al., 2016; Li et al., 2019) .

A fourth possibility to improve model performance is to address data uncertainty of streamflow observations and of precipitation input. Errors in streamflow measurements caused by stage-discharge rating-curve uncertainty (Coxon et al., 2015; Kiang et al., 2018) influence model calibration and evaluation. To improve uncertainty estimates, such uncertainty should be accounted for by explicitly considering streamflow measurement uncertainty in model calibration (McMillan et al., 2010). In addition, the uncertainty of the precipitation product used to drive a hydrological model can lead to differences in observed and simulated flows (Te Linde et al., 2007; Renard et al., 2011). 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 (Haylock et al., 2008; Risser et al., 2019). Such spatial uncertainty could be accounted for by using probabilistic analyses of precipitation fields (Newman et al., 2015a; Frei and Isotta, 2019). The consideration of such input uncertainty is particularly important if we are interested in future changes because of climate model and scenario uncertainty, where precipitation uncertainty is specifically pronounced (Chen et al., 2014; Lopez-Cantu et al., 2020). Even though many of these possibilities have been discussed in previous studies, their consideration in flood analyzes is not a standard practice.

5 Conclusions

Our model comparison shows that flood characteristics are not always well captured in hydrological models developed for research studies – even when the models have been calibrated with a calibration metric perceived suitable for flood modeling, the Kling–Gupta efficiency metric (E_{KG}). 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. 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' according to the 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.

The limited capability of the models in reproducing local and spatial flood characteristics and the sensitivity of runoff to precipitation inputs is partly attributed to model structure and partly to a reliance of the calibration on an individual variable (streamflow) and 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. 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. This study underscores the importance of improving the representation of magnitude, timing, spatial connectedness, and flood generating processes. Potential ways of achieving such improvements include developing flood-focused, multi-objective and spatial calibration metrics, improving flood generating process representations through model structure comparisons, and reducing uncertainty in precipitation input. Such steps are recommended to improve the reliability of flood simulations and ultimately local and regional flood hazard assessments 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 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

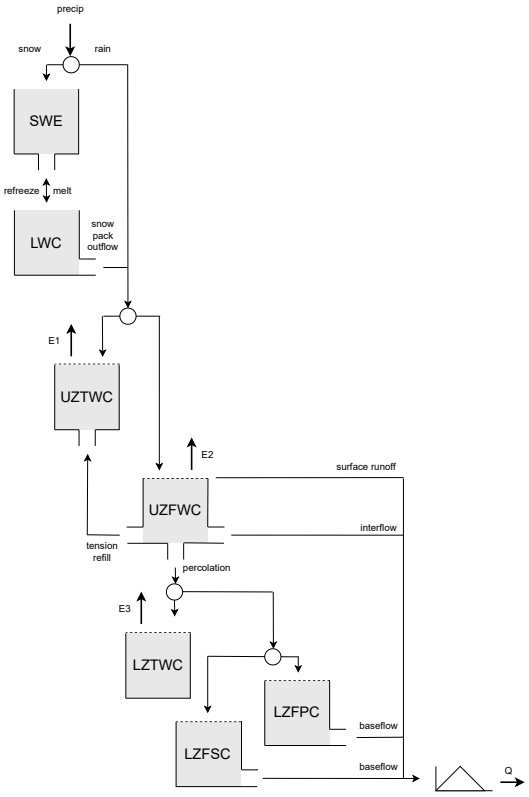

Figure A1. Structure of the Snow17/SAC-SMA model. Fluxes: precipitation (precip), snow, rain, snowmelt (melt), refreeze, snowpack outflow, evapotranspiration (E1, E2, and E3), tension refill, surface runoff, interflow percolation, baseflow, simulated discharge (Q). States: snow-water-equivalent (SWE),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

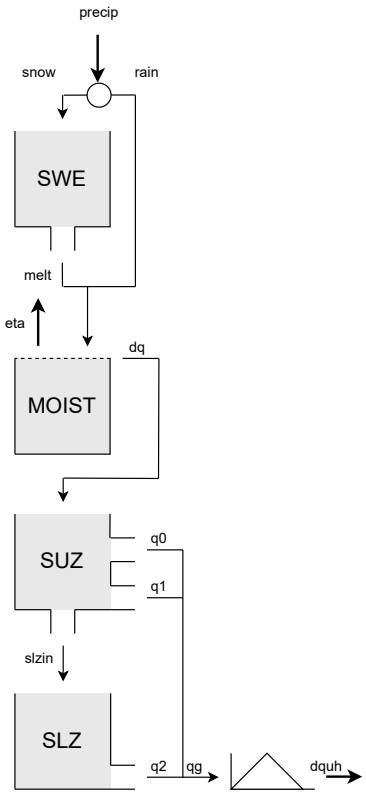

Figure A2. Structure of the TUW HBV model. Fluxes: precipitation (precip), snow, rain, snowmelt (melt), actual evapotranspiration (eta), runoff (dq), surface runoff (q0), subsurface runoff (q1), baseflow (q2), simulated runoff (qg), simulated discharge (dquh), input from upper to lower storage (slzin).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

A3 VIC

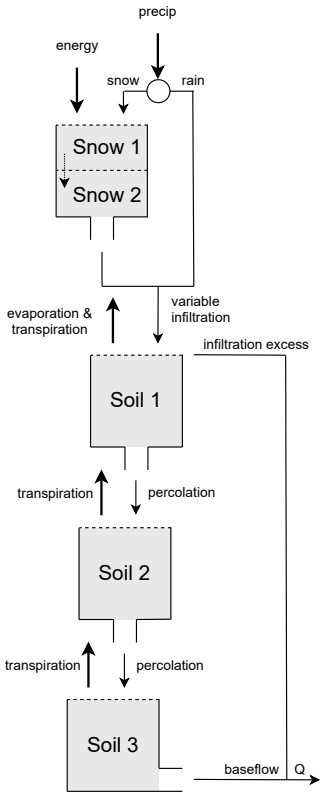

Figure A3. Fluxes: precipitation (precip), energy, snow, rain, variable infiltration, evaporation and transpiration, infiltration excess, baseflow, percolation, transpiration, simulated runoff (Q). Storage: snow layer 1 (Snow 1), snow layer 2 (Snow 2), soil layer 1 (Soil 1), soil layer 2 (Soil 2), soil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

A4 mHM

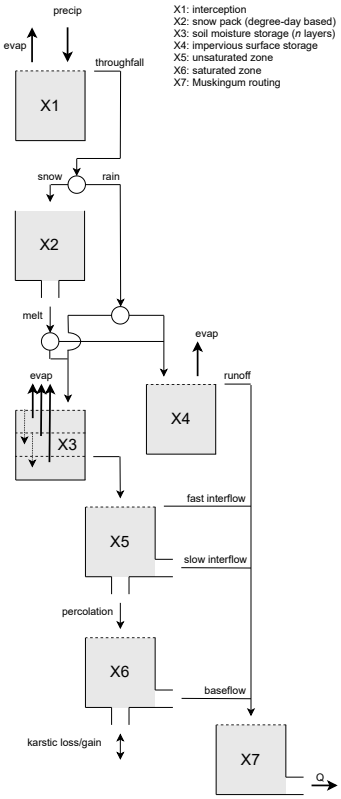

Figure A4. Fluxes: precipitation (precip), evapotranspiration (evap), throughfall, snow, rain, snowmelt (melt), runoff, fast interflow, slow interflow, percolation, baseflow, karstic loss/gain, simulated discharge (Q). Storage: Interception storage (X1), snow pack (X2), soil moisture storage (X3), impervious 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 editor and the four 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, 24, 3951–3966, https://doi.org/10.5194/hess-24-3951-2020, 2020b.

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

Burnash, R. J. C., 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.

Cooley, D., Cisewski, J., Erhardt, R. J., Jeon, S., Mannshardt, E., Omolo, B. O., and Sun, Y.: A survey of spatial extremes: Measuring spatial dependence and modeling spatial effects, Revstat Statistical Journal, 10, 135–165, 2012.

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.

Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R., and Smith, P.: A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations, Water Resources Research, 51, 5531–5546, https://doi.org/10.1002/2014WR016532, 2015.

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, 56, e2019WR026 085, 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.

Efstratiadis, A. and Koutsoyiannis, D.: One decade of multi-objective calibration approaches in hydrological modelling: a review, Hydrological Sciences Journal, 55, 58–78, https://doi.org/10.1080/02626660903526292, 2010.

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.

Frei, C. and Isotta, F. A.: Ensemble spatial precipitation analysis from rain gauge data: Methodology and application in the European Alps, Journal of Geophysical Research: Atmospheres, 124, 5757–5778, https://doi.org/10.1029/2018JD030004, https://doi.org/10.1029/2018JD030004, 2019.

Gilleland, E.: Testing competing precipitation forecasts accurately and efficiently: The Spatial Prediction Comparison Test, Monthly Weather Review, 141, 340–355, https://doi.org/10.1175/MWR-D-12-00155.1, 2013.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resources Research, 34, 751–763, https://doi.org/10.1029/97WR03495, 1998.

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.

Haylock, M. R., Hofstra, N., Klein Tank, A. M., Klok, E. J., Jones, P. D., and New, M.: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950-2006, Journal of Geophysical Research Atmospheres, 113, https://doi.org/10.1029/2008JD010201, 2008.

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.

Hrachowitz, M. and Clark, M. P.: HESS Opinions: The complementary merits of competing modelling philosophies in hydrology, Hydrology and Earth System Sciences, 21, 3953–3973, https://doi.org/10.5194/hess-21-3953-2017, 2017.

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, W04 512, 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.

Kempen, G. V., Wiel, K. V. D., and Melsen, L. A.: The impact of hydrological model structure on the simulation of extreme runoff events, Natural Hazards and Earth System Sciences, p. under review, https://doi.org/10.5194/nhess-2020-154, 2020.

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.

Kiang, J. E., Gazoorian, C., McMillan, H., Coxon, G., Le Coz, J., Westerberg, I. K., Belleville, A., Sevrez, D., Sikorska, A. E., Petersen-Øverleir, A., Reitan, T., Freer, J., Renard, B., Mansanarez, V., and Mason, R.: A comparison of methods for streamflow uncertainty estimation, Water Resources Research, 54, 7149–7176, https://doi.org/10.1029/2018WR022708, 2018.

Klemes, V.: Operational testing of hydrological simulation models, Hydrological Sciences Journal, 31, 13–24, https://doi.org/10.1080/02626668609491024, 1986.

Knoben, W. J. M., 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.

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., and Woods, R. A.: A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments, Water Resources Research, 56, e2019WR025 975, https://doi.org/10.1029/2019WR025975, https://doi.org/10.1029/2019WR025975, 2020.

Koch, J., Jensen, K. H., and Stisen, S.: Toward a true spatial model evaluation in distributed hydrological modeling: Kappa statistics, Fuzzy theory, and EOF-analysis benchmarked by the human perception and evaluated against a modeling case study, Water Resources Research, 51, 1225–1246, https://doi.org/10.1002/2014WR016607, 2015.

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.

Kraußе, T., Cullmann, J., Saile, P., and Schmitz, G. H.: Robust multi-objective calibration strategies-possibilities for improving flood forecasting, Hydrology and Earth System Sciences, 16, 3579–3606, https://doi.org/10.5194/hess-16-3579-2012, 2012.

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.

Li, D., Lettenmaier, D. P., Margulis, S. A., and Andreadis, K.: The role of rain-on-snow in flooding over the conterminous United States, Water Resources Research, 55, 8492–8513, https://doi.org/10.1029/2019WR024950, 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.

Madsen, H.: Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives, Advances in Water Resources, 26, 205–216, https://doi.org/10.1016/S0309-1708(02)00092-1, 2003.

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., 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.

Melsen, L. A. 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.

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., Craig, J., Nijssen, B., Wood, A., Gutmann, E., Mizukami, N., Brekke, L., and Arnold, J. R.: Gridded ensemble precipitation and temperature estimates for the Contiguous United States, Journal of Hydrometeorology, 16, 2481–2500, https://doi.org/10.1175/jhm-d-15-0026.1, 2015a.

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 characteristics 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, 2015b.

Pomeroy, J. W., Fang, X., and Marks, D. G.: The cold rain-on-snow event of June 2013 in the Canadian Rockies — characteristics and diagnosis, Hydrological Processes, 30, 2899–2914, https://doi.org/10.1002/hyp.10905, 2016.

Pool, S., Vis, M. J. P., 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.

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., and Franks, S. W.: Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation, Water Resources Research, 47, W11 516, https://doi.org/10.1029/2011WR010643, 2011.

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.

Seibert, J.: Reliability of model predictions outside calibration conditions, Nordic Hydrology, 34, 477–492, 2003.

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.

Spieler, D., Mai, J., Craig, J. R., Tolson, B. A., and Schütze, N.: Automatic model structure identification for conceptual hydrologic models, Water Resources Research, https://doi.org/10.1029/2019wr027009, 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.: On the need to test hydrological models under changing conditions, Hydrological Sciences Journal, 60, 1165–1173, https://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.