Evaluating the suitability of hydrological models for flood impact assessments

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

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

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

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

1 Introduction

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

A suitable model ideally reproduces different aspects of flooding, including local characteristics such as event magnitude and timing. It has been shown, however, that capturing magnitude and timing is challenging when standard calibration metrics are used individually for parameter estimation (Lane et al., 2019; Brunner and Sikorska, 2018; Mizukami et al., 2019). For example, one widely-used metric that is considered integrative compared to others (e.g., bias, correlation) is the Nash–Sutcliffe efficiency (E_{NS}; Nash and Sutcliffe 1970), where the sum-of-squares error metric focuses attention on high flows. However, E_{NS} is formulated so that its optimal value actually underestimates flow variability (Gupta et al., 2009). Using a related metric, the Kling–Gupta efficiency (E_{KG}; Gupta et al. 2009) can partially overcome this deficiency and improve simulations of peak flows (Mizukami et al., 2019). Yet to achieve further improvement, a broader range of application-specific evaluation metrics is typically required, including objectives that directly characterize hydrologic phenomena (or 'signatures') such as peak flows, flood volumes and timing, recession rates, and seasonal hydrograph shape. Considering multiple objectives in a step-wise calibration sequence, either manual or automated, is common in agencies that implement models for applications such as flood forecasting (Hogue et al., 2000), and strengthens their ability to provide reliable flood predictions. The use of multiple objectives, however, can lead to a decrease in performance with respect to any individual flow signature not considered as an objective (Mizukami et al., 2019). Despite their deficiencies with respect to extremes, individual 'integrative' standard calibration metrics such as E_{NS} or E_{KG} are often used in research modeling studies, even when the focus is on floods and their future changes (for E_{NS} based calibration see e.g. Hundecha and Merz 2012; Köplin et al. 2014; Vormoor et al. 2015; Wobus et al. 2017 and for E_{KG} based calibration see e.g. Harrigan et al. 2020; Hirpa et al. 2018; Huang et al. 2018; Thober et al. 2018; Brunner and Sikorska 2018).

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

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

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

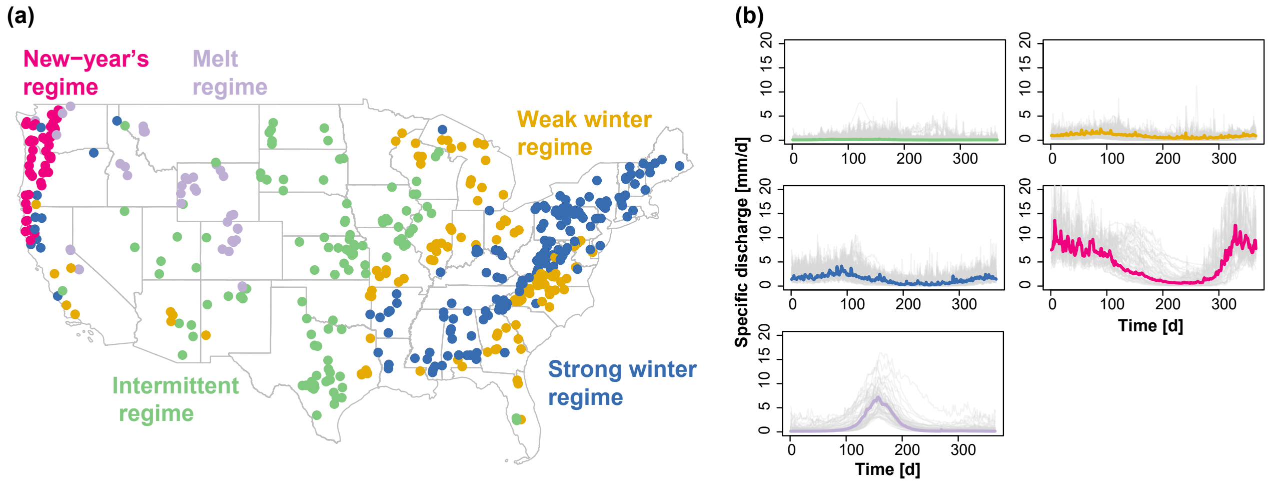

Figure 1. a) Map of the 488 catchments in the conterminous United States belonging to the five regime classes indicated by their gauge location: 1) Intermittent, 2) weak winter, 3) strong winter, 4) New Year's, and 5) melt. b) Median regime per regime class (colored lines) and variability of regimes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

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

2 Data and Methods

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

2.1 Model simulations

We use daily streamflow simulations for the period 1981-2008 generated with four well-known hydrological models (Addor and Melsen, 2019) offering different model structures and complexity: the lumped SAC model (Figure A1; Burnash et al., 1973), the lumped HBV model (Figure A2; Bergström, 1976), the lumped version of the VIC model (Figure A3; Liang et al., 1994), and the grid-based, distributed mesoscale hydrologic model mHM (Figure A4; Kumar et al., 2013; Samaniego et al., 2010). The model parameters were calibrated on streamflow observations by minimizing E_{KG} by Melsen et al. (2018) using Sobol-based Latin hypercube sampling (Bratley and Fox, 1988) for SAC, HBV, and VIC and by Mizukami et al. (2019) for mHM using multi-scale parameter regionalization where the transfer function parameters were identified using the dynamically dimensioned search algorithm (Tolson and Shoemaker, 2007). E_{KG} is defined as:

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

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

For SAC, Melsen et al. (2018) calibrated and evaluated 18 out of the 35 parameters available in the coupled Snow-17 and SAC-SMA modeling system, for HBV 15 parameters, for VIC 17 parameters, and for mHM Rakovec et al. (2019) and Mizukami et al. (2019) calibrated and evaluated up to 48 parameters. All the models were driven with daily, spatially lumped meteorological forcing data representing current climate conditions: SAC, HBV, and VIC were driven with Daymet meteorological forcing (1km resolution; Thornton et al., 2012) and mHM with the forcing by Maurer et al. (2002) (12km resolution) both derived from observed precipitation and temperature. SAC, HBV, and VIC were calibrated and evaluated on the period 1985–2008 while mHM was calibrated on the period 1999–2008 and evaluated on the period 1989-1999. After calibration, all four models were run for the period 1980–2008 (calendar years), where the period 1980-1981 was here used for spin-up and therefore discarded from the analysis.

Model performance in terms of E_{KG} varies spatially and is related to the hydrological regime (Figure 2). It is overall lowest for catchments with intermittent regimes and a weak seasonality and highest for catchments with a strong seasonality such as a melt and New Year's regime. However, there is a high within-class variability in model performance. The finding that intermittent regimes are challenging to model successfully is well known in hydrology and reproduced in many studies, e.g., Unduche et al. (2018), who show that hydrological modeling on Prairie watersheds is very complex (Hay et al., 2018). Intermittent regimes may suffer in calibration if they rely solely on correlation-type measures because their day to day variation is more difficult to reproduce than a more pronounced and regular seasonality. Overall model performance decreases from mHM (median E_{KG} 0.69), over SAC (median E_{KG} 0.63) and VIC (median E_{KG} 0.60) to HBV (median E_{KG} 0.52). In addition to streamflow, we use areal precipitation and simulated soil moisture to explain potential differences in model performance.

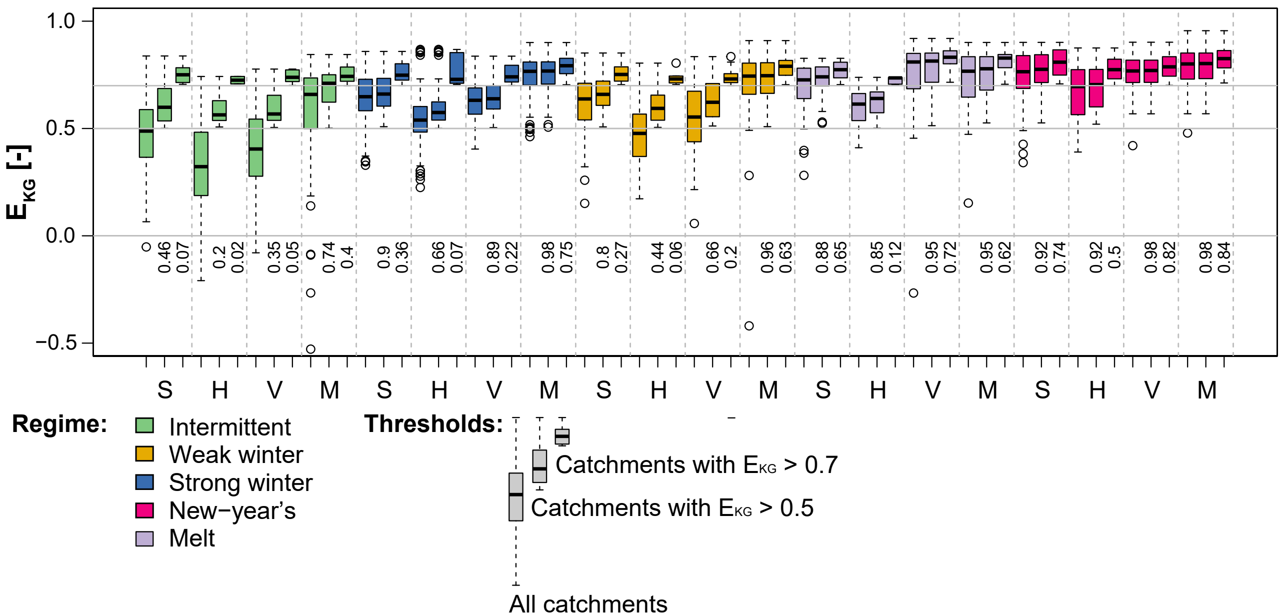

Figure 2. Model performance in terms of E_{KG} over the period 1981–2008 for the four models SAC (S), HBV (H), VIC (V), and mHM (M) per hydrological regime: intermittent (114 catchments), weak winter (108), strong winter (176), New Year's (50), and melt (40). For each model and regime, three boxplots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

2.2 Model evaluation for floods

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

2.2.1 Flood event identification

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

2.2.2 Flood characteristics at individual sites

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

2.2.3 Spatial flood dependence

We then use the data sets resulting from Step 2 to evaluate how models reproduce overall and seasonal spatial flood dependence. To do so, we use the connectedness measure introduced by Brunner et al. (2020a), which quantifies the number of catchments with which a specific catchment co-experiences floods. The number of concurrent flood events for a pair of stations is determined based on a data set consisting of the dates of flood occurrences across all catchments. This set is converted into a binary matrix which specifies for each catchment whether or not it is affected by a certain event. The matrix compiled using observed streamflow time series contained 1164 events among which 258 occur in winter, 291 in spring, 324 in summer, and 291 in fall. Following the definition used by Brunner et al. (2020a), a catchment is connected to another catchment if they share a certain number of events. We here used an event threshold of 1% of the total or seasonal number of events to define connectedness (all months: 12 events, seasons: 3 events). We computed actual errors in flood connectedness by subtracting observed from simulated connectedness over all seasons and per season.

2.2.4 Flood triggers

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

2.2.5 Floods under change

In addition to assessing model performance under current climate conditions, we would like to understand potential, additional challenges arising when interested in future conditions. To do so, we look at how models translate changes in event temperature and precipitation into changes in POT discharge by performing a resampling-based sensitivity analysis. This sensitivity analysis aims at evaluating whether a model is still reliable under climate conditions different from the ones used in model calibration similar to split-sample or differential split-sample calibration/validation schemes (Coron et al., 2012; Refsgaard et al., 2014; Thirel et al., 2015). To perform this sensitivity analysis, we generate surrogate time series of temperature, precipitation, and streamflow for each catchment (Wood et al., 2004; Brunner et al., 2020b). To generate these series, we randomly sample a series of years with replacement in the period 1981–2008 which we use to compose time series consisting of the daily values corresponding to these years for each of the three variables. For each of the surrogate series, we again extract POT flood events using the same procedure as described under Step 1. For each of the extracted events we then determine temperature and precipitation. We use the sets of peak discharge, event temperature and event precipitation to compute mean event discharge, temperature, and precipitation, which enables the derivation of a relationship between mean POT discharge and the two meteorological variables during events. We repeat the resampling $n = 500$ times to derive a relationship between changes in mean event temperature and precipitation and changes in mean POT streamflow. This resampling experiment results in a response surface of POT discharge spanned by mean event temperature and mean event precipitation for each catchment. We summarize the results obtained at individual locations by computing horizontal and vertical sensitivity gradients on these reaction surfaces using a linear regression model. The horizontal gradient describes the strength of POT discharge changes in response to event temperature changes while the vertical gradient describes the strength of change in response to changes in event precipitation. Conducting this experiment for both observed and simulated time series allows for the determination of whether the models react to changes in mean event temperature and precipitation in the same way as the real world system and are therefore suitable for the use in climate change impact assessments on floods. If models produce different climate sensitivities than the ones seen in the observations, the use of models to simulate sets of flood events for future conditions may preclude reliable change assessments.

3 Results

3.1 Flood characteristics at individual sites

Model performance at individual sites with respect to the number of events, event magnitude, and timing varies by model and hydrological regime type (Figure 3). For most catchments, the median deviation between the simulated and observed number of flood events lies close to zero (SAC: -3 events, HBV: -1, VIC: -1, mHM: 0). However, the simulations result in over- and underestimations of the number of events depending on the catchment (1st and 3rd quartiles for SAC: -9, 4; HBV: -8, 15; VIC: -7, 6; mHM: -6, 6). The overestimation is strongest for HBV, which overestimates the number of events for catchments with intermittent, weak winter, and melt regimes (Brunner et al., 2020b). The overestimation of the number of events by HBV may be explained by its fast response to precipitation as expressed through its model parameter β, which introduces non-linearity to the system (Viglione and Parajka, 2020). Event magnitude in terms of peak discharge is generally underestimated for all regime types independent of the model. Underestimation is in line with previous studies showing that using E_{KG} individually results in an underestimation of peak flow (Mizukami et al., 2019) due to an underestimation of variability, which will result in an under-representation of extremes (Katz and Brown, 1992). Another factor potentially contributing to this underestimation is that the models were forced with spatially lumped instead of distributed data, which may smooth the simulated discharge response.

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

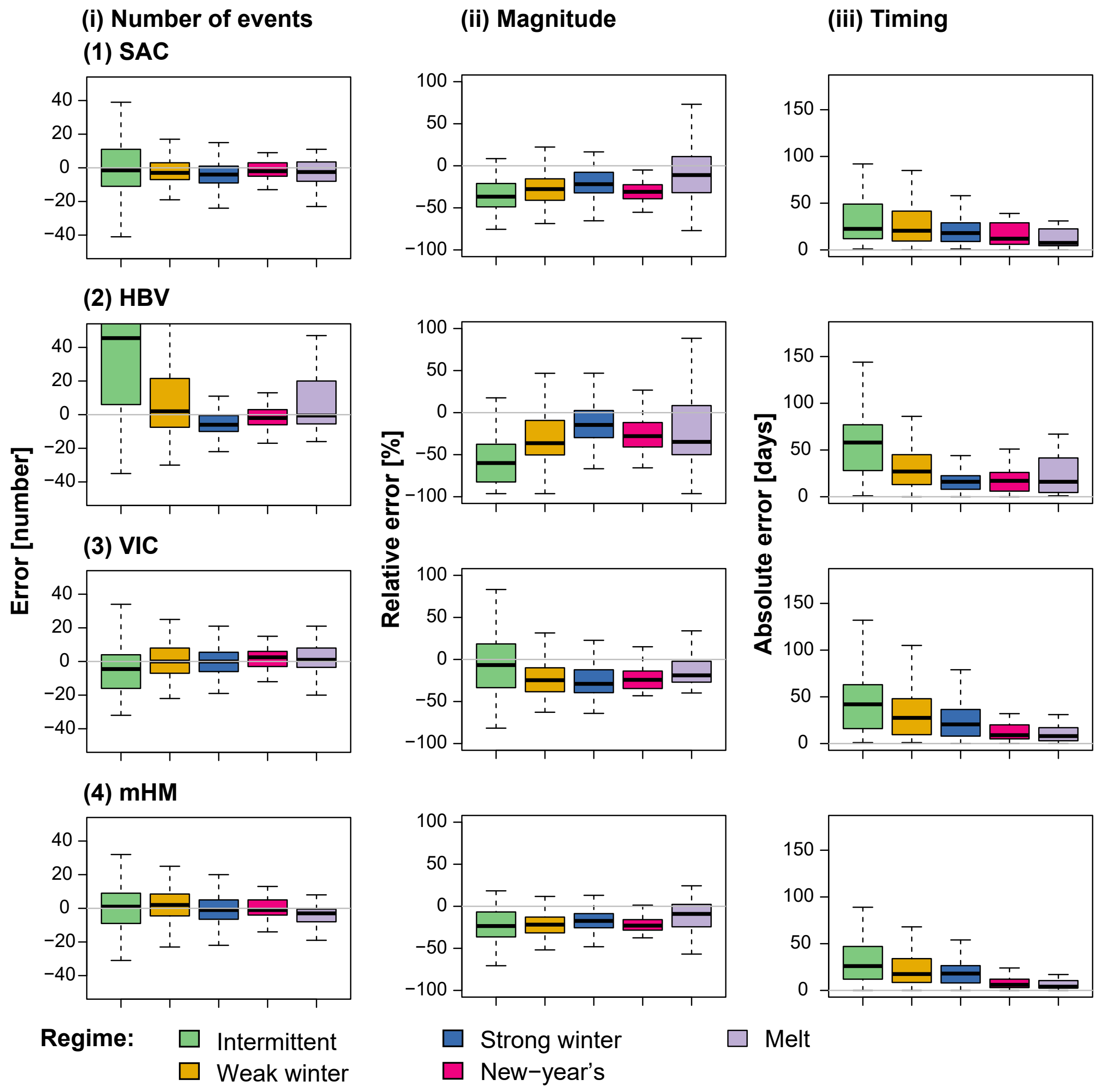

Figure 3. Model errors per regime type computed over the period 1981–2008: intermittent (114 catchments), weak winter (108), strong winter (176), New Year's (50), and melt (40) (Figure 1). Errors are shown for (i) number of events (error in number of events), (ii) magnitude (mean relative error), and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

a few months (Brunner et al., 2020a). Over all, there is no clear tendency of one model to perform better than the other ones. However, there are slight differences in model performance which suggests that a 'most suitable model' could be identified for a specific application at hand, where a certain region or variable is of interest.

3.2 Spatial flood dependencies

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

3.3 Flood triggers

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

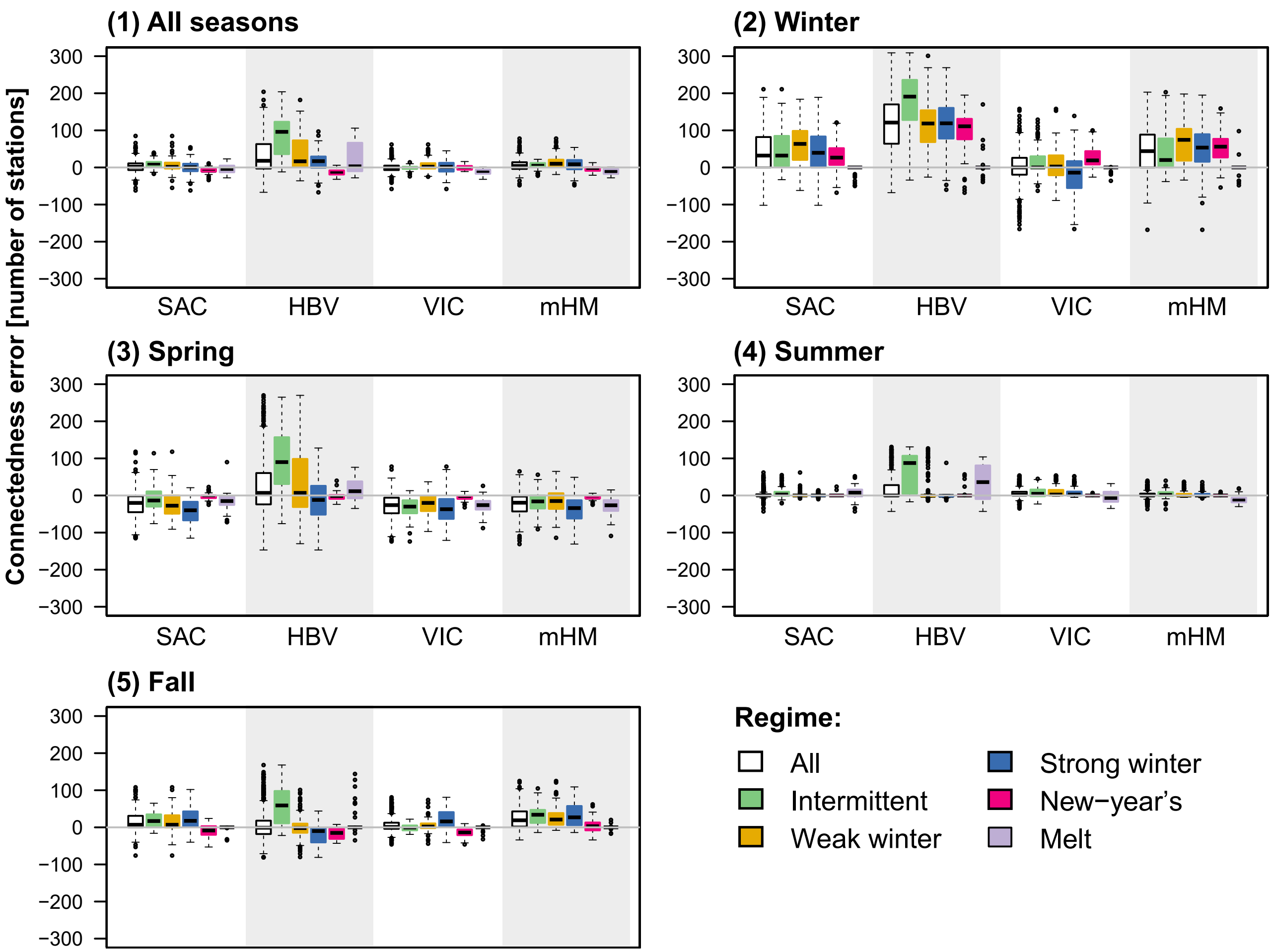

Figure 4. Overall (1) and seasonal (2–5) errors in flood connectedness (simulated minus observed connectedness), i.e. number of catchments a catchment is sharing at least 1% of the total number of flood events with, for the four models SAC, HBV, VIC, and mHM over all regimes and per regime: intermittent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

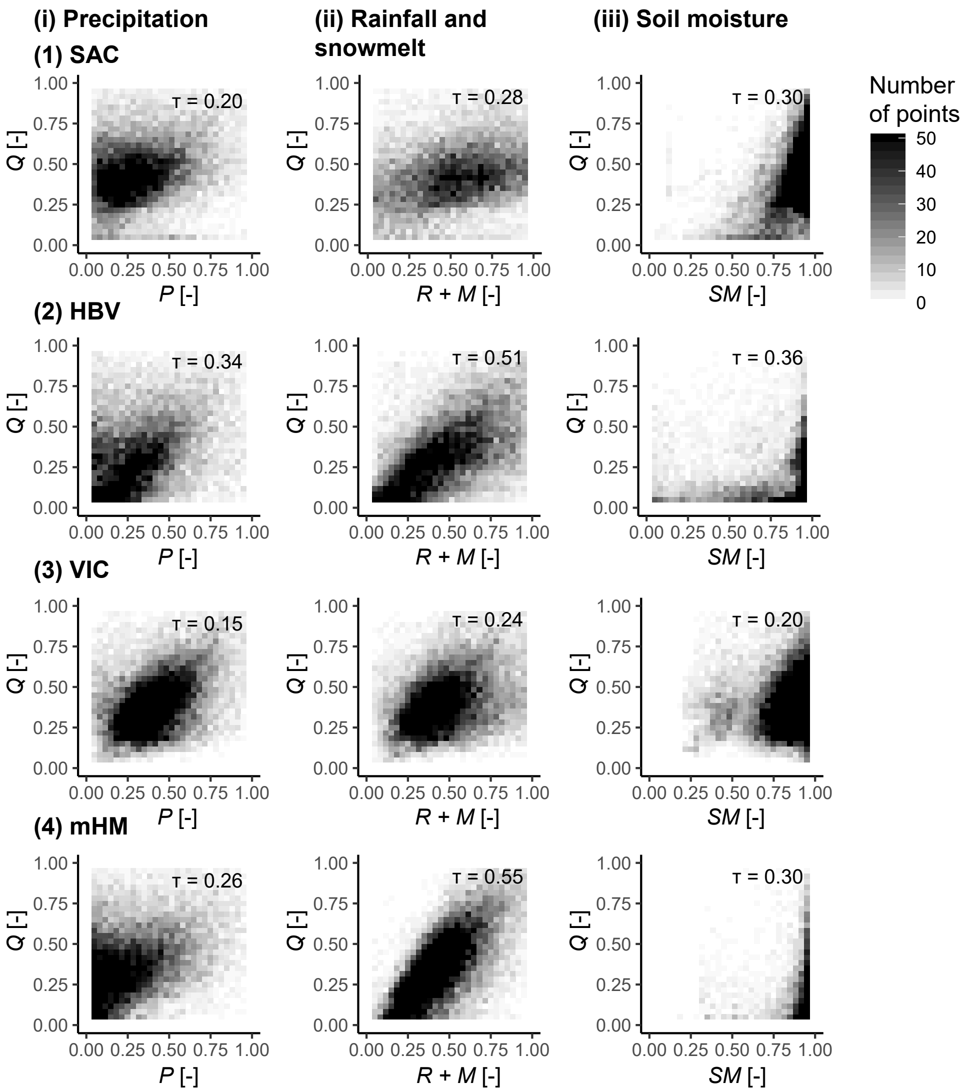

Figure 5. Simulated relationships between normalized flood discharge (Q) and normalized precipitation (i, P), rainfall and snowmelt (ii, $R+M$), and soil moisture (iii, SM, upper two soil layers for mHM) over all catchments represented by a binned scatter plot for the four hydrologic models (1) SAC,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

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

We here show that model performance to some degree depends on model choice and that many different combinations of forcing and model states can simulate floods. In addition, model performance may depend on the uncertainty of streamflow observations (McMillan et al., 2010) used for calibrating and evaluating the model or on input uncertainty, i.e. the precipitation product used to drive the models (Te Linde et al., 2007). Precipitation products may show observation uncertainties (Mcmillan et al., 2012) and underestimate extreme rainfall or the spatial dependence of extreme precipitation at different locations because spatial smoothing or averaging during the gridding process reduces variability (Risser et al., 2019). The importance of input uncertainty is particularly pronounced if we are interested in future changes because of climate model and scenario uncertainty (Chen et al., 2014; Lopez-Cantu et al., 2020).

3.4 Floods under change

In addition to looking at how well local and spatial flood characteristics are represented by models, we look at how changes in temperature and event precipitation are translated into changes in flood flows to assess each models' suitability for climate impact assessments on floods. Our sensitivity analysis shows that the models have difficulty translating changes in event temperature and precipitation into sensitivities of flood flows (Figure 6), which can be problematic if we would like to use such models in climate change assessments. Generally, flood flows show a relatively low sensitivity to changes in mean event precipitation and temperature. This is in contrast to the behavior for mean flow, which is strongly influenced by changes in mean precipitation as demonstrated in a similar experiment by Brunner et al. (2020b). The much stronger relationship between mean precipitation and flow than between event precipitation and flow might arise because mean flow is a climate signal (Knoben et al., 2018), whereas floods are more an event (higher frequency, short-term) signal. However, some catchments, e.g. the Tucca Creek (New Year's regime) show a clear relationship between peak magnitude and both event temperature and precipitation. While these relationships are captured for some catchments (e.g. Blackwater River, weak winter regime or Tucca Creek, New Year's regime), they aren't in other catchments. The simulated sensitivities may even point in another direction than the observed ones (e.g. Pacific Creek, melt regime). In the case of melt regimes, the misrepresentation of flood sensitivities by models suggests that they may have difficulty simulating snow-influenced flooding.

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

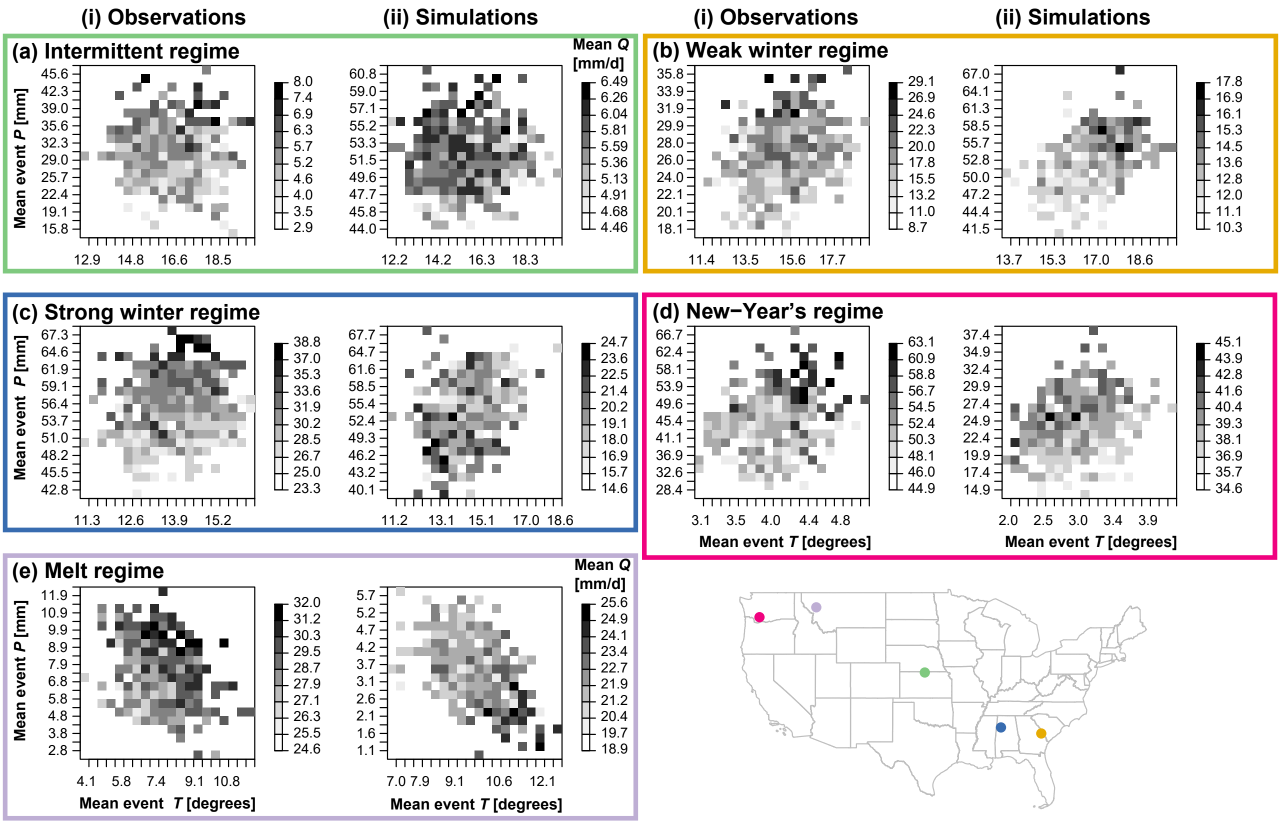

Figure 6. Climate sensitivity analysis for the VIC model: Dependence of mean POT magnitude (Q) on mean flood event precipitation (1-day; P) and mean flood temperature (T) for five example catchments, those with the best E_{KG} per regime type: intermittent regime (green; USGS ID 09210500 Fontanelle Creek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

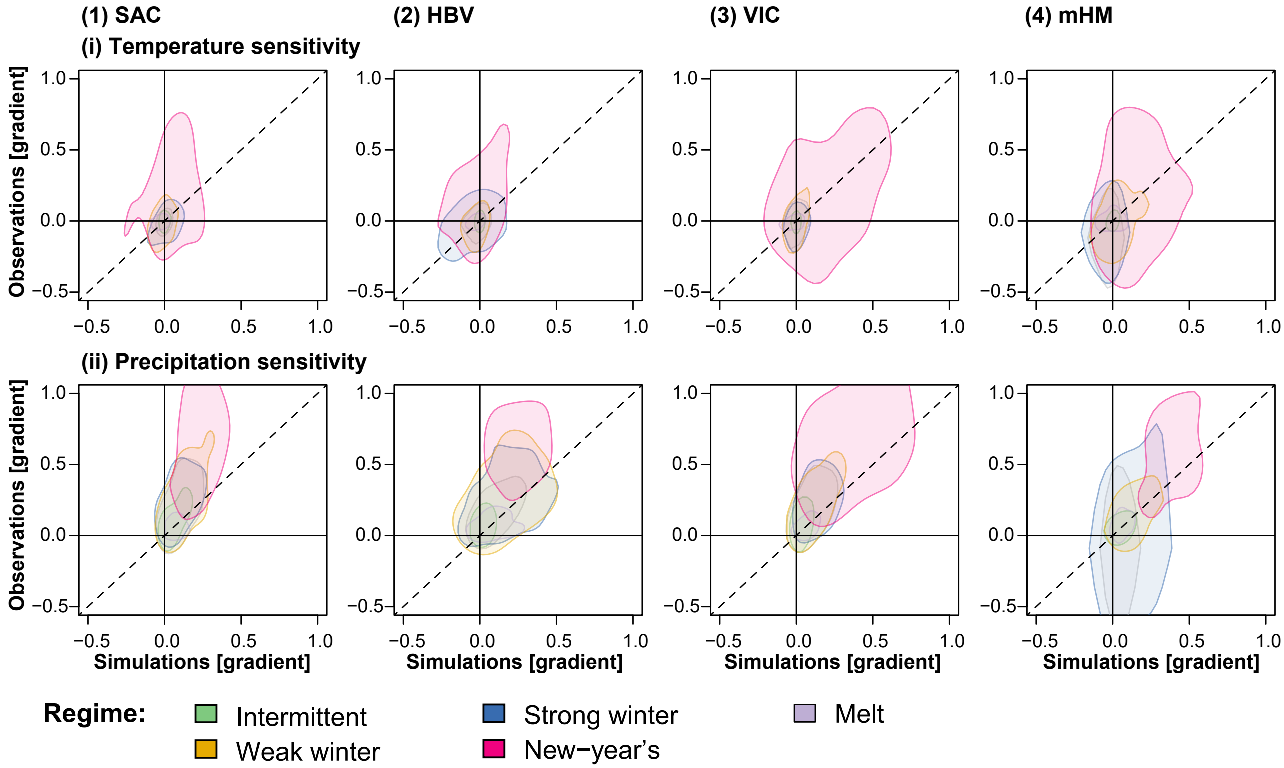

Figure 7. Observed vs. simulated (i) horizontal (temperature) and (ii) vertical (precipitation) climate sensitivities for floods represented by two-dimensional kernel density estimates for the four models (1) SAC, (2) HBV, (3) VIC, and (4) mHM for the five regime types: intermittent (114 catchments), weak winter (108), strong winter (176), New Year's (50), and melt (40) (Figure 1). Positive and negative values indicate positive and negative associations of precipitation and temperature with peak flow, respectively. Values on the dashed line indicate correspondence between observed and modeled sensitivity gradients.

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

The results of this study indicate that the hydrological models used in this study have limited capability in reproducing observed hydrologic sensitivities during flooding. These limitations may be related to input uncertainties (Te Linde et al., 2007), equifinality in process contributions for simulations with (very) similar efficiency scores, leading to an inability to unambiguously identify the appropriate relative process contributions (Khatami et al., 2019) or insufficient model calibration (Fowler et al., 2016). We illustrate that reliance on an individual calibration metric (E_{KG}) can lead to simulation performance deficits for phenomena of interest, including an underestimation of streamflow variability (Mizukami et al., 2019) and peak flood magnitudes and a misrepresentation of timing. As is evident in some existing practice-oriented applications of hydrological models (Hogue et al., 2000; Unduche et al., 2018; World Meteorological Organization, 2011), the simulation of floods and other hydrologic phenomena is likely to be improved by using more tailored model calibration strategies. The former would include either giving more weight to the variability component of an integrative metric such as the E_{KG} (Pool et al., 2017; Mizukami et al., 2019); whereas the latter might include optimizing explicitly for key flood characteristics (e.g., peak flow, volume, timing) and/or metrics depicting the fidelity of the model representation of soil moisture and snowmelt, within a multi-objective model calibration process (Moussa and Chahinian, 2009; Sikorska et al., 2018; Sikorska-Senoner et al., 2020). The spatial representation of extremes may also be improved by considering spatially distributed features of model response within a spatial calibration framework (Dembélé et al., 2020; Koch et al., 2018).

4 Conclusions

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

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

We conclude that calibration using only an individual model performance metric or variable can result in model implementations that have limited value for specific model applications, such as local and in particular spatial flood hazard analyses and change impact assessments. Despite its shortcomings, this practice has become increasingly more common and accepted in the research literature. Yet, our analysis illustrates that the development and adoption of more comprehensive multi-objective and multi-variable calibration strategies are needed to significantly improve model performance regarding floods under both current and future climate conditions.

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

Appendix A: Model illustrations

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

- Snow17/SAC-SMA: analysis of the model's description (National Weather Service NOAA, 2002): https://www.nws.noaa.gov/oh/hrl/general/chps/Models/Sacramento_Soil_Moisture_Accounting.pdf and source code.
- TUW HBV: analysis of the model's source code (Viglione and Parajka, 2020).
- VIC: descriptions of VIC in Melsen et al. (2018); Melsen and Guse (2019) and on analysis of the v4.1.2h source code (https://github.com/UW-Hydro/VIC/releases/tag/VIC.4.1.2.h).
- mHM: analysis of the model's source code (https://git.ufz.de/mhm/mhm/-/tree/5.7) and a diagram provided in (Kumar et al., 2010).

A1 Snow17/SAC-SMA

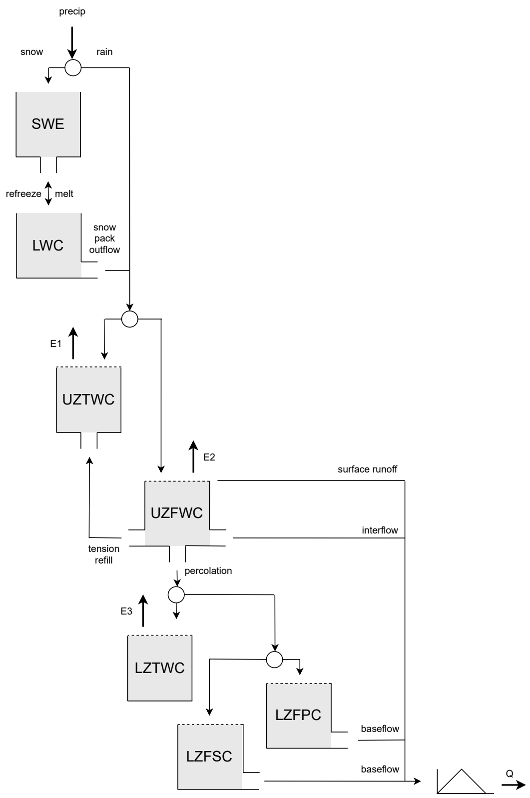

Figure A1. Structure of the Snow17/SAC-SMA model. Fluxes: precipitation (precip), snow, rain, snowmelt (melt), refreeze, snowpack outflow, evapotranspiration (E1, E2, and E3), tension refill, surface runoff, interflow percolation, baseflow, simulated discharge (Q). States: snow-water-equivalent (SWE),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

A2 TUW-HBV

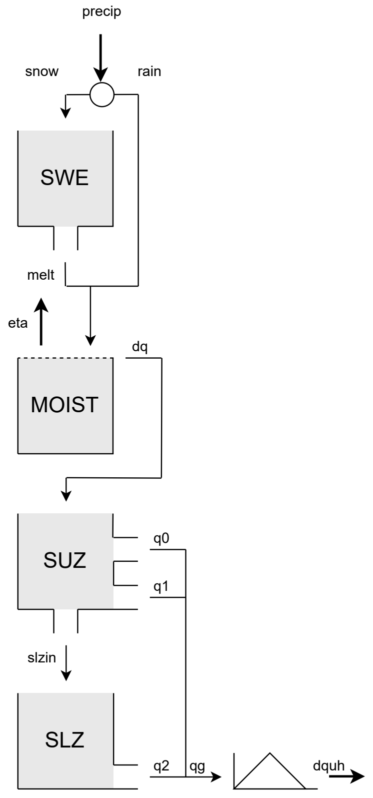

Figure A2. Structure of the TUW HBV model. Fluxes: precipitation (precip), snow, rain, snowmelt (melt), actual evapotranspiration (eta), runoff (dq), surface runoff (q0), subsurface runoff (q1), baseflow (q2), simulated runoff (qg), simulated discharge (dquh), input from upper to lower storage (slzin).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

A3 VIC

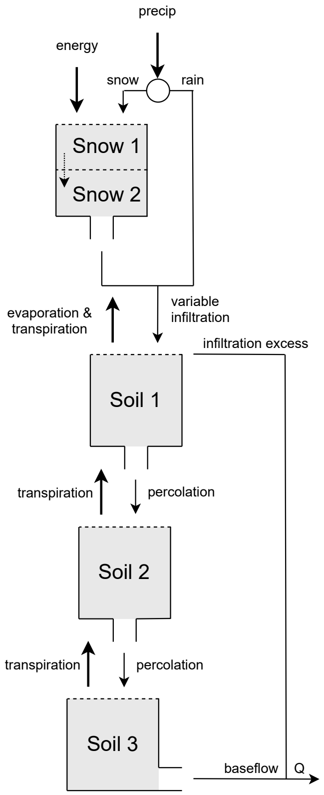

Figure A3. Fluxes: precipitation (precip), energy, snow, rain, variable infiltration, evaporation and transpiration, infiltration excess, baseflow, percolation, transpiration, simulated runoff (Q). Storage: snow layer 1 (Snow 1), snow layer 2 (Snow 2), soil layer 1 (Soil 1), soil layer 2 (Soil 2), soil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

A4 mHM

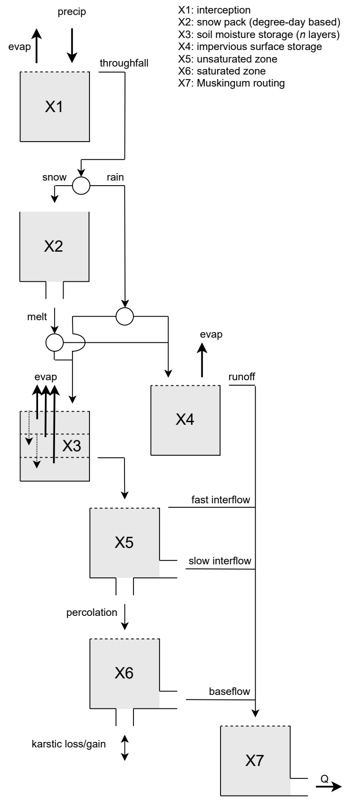

Figure A4. Fluxes: precipitation (precip), evapotranspiration (evap), throughfall, snow, rain, snowmelt (melt), runoff, fast interflow, slow interflow, percolation, baseflow, karstic loss/gain, simulated discharge (Q). Storage: Interception storage (X1), snow pack (X2), soil moisture storage (X3), impervious 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 three reviewers for their constructive feedback, which helped to reframe and clarify the storyline.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Huang, S., Kumar, R., Rakovec, O., Aich, V., Wang, X., Samaniego, L., Liersch, S., and Krysanova, V.: Multimodel assessment of flood characteristics in four large river basins at global warming of 1.5, 2.0 and 3.0 K above the pre-industrial level, Environmental Research Letters, 13, 124 005, https://doi.org/10.1088/1748-9326/aae94b, 2018.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set 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, 2015.

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

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

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

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

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

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

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

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

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

Te Linde, A. H., Aerts, J., Dolman, H., and Hurkmans, R.: Comparing model performance of the HBV and VIC models in the Rhine basin, in: Quantification and Reduction of Predictive Uncertainty for Sustainable Water Resources Management, 313, pp. 278–285, 2007.

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

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

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

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

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

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

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

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

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

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

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

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