Replies to the comments of Anonymous Referee #1

Responses are provided in blue and proposed revision are in Red. Original reviewer comments are in black. Line and page numbers refer to the original manuscript.

Based upon comments from all three reviewers, we have revisited our model calibration procedure and have been able to obtain acceptable performance from the snowpack routine. This involved a "2-step" calibration process in which warm season processes are calibrated first, and then "warm season parameters " are held constant during subsequent calibration of snowpack-related parameters. This recalibration of HBV is done using both CPC and Stage IV rainfall. We have also added a section on model validation to the revised manuscript, again based on comments from all three reviewers requesting additional validation results. Since all three reviewers provided critiques on these topics, we discuss these two changes before addressing specific comments from individual reviewers.

We have revised model calibration part in the original manuscript, P9, line 15-24, to:

We calibrated the HBV models using both CPC and Stage IV rainfall, and most parameters are the same for CPC- and Stage IV-based models, except for three snow routine parameters (TT, CFMAX, SFCF) and three recession coefficients (K0, K1, K2), allowing for the variability of model parameters for different climate conditions. For each model setup, we first calibrated the model with snowpack routine "turned off" (by setting TT parameter to a very low value) to obtain parameters that can simulate summer floods adequately. Then, keeping these optimized non-snow routine parameters unchanged, we calibrated the snow routine parameters.

To determine the optimized model parameter sets in each procedures, we followed the Genetic Algorithm and Powell (GAP) optimization method as presented by Seibert (2000), which is briefly summarized here. First, 5000 parameter sets are randomly generated from a uniform distribution of the values of each parameter (Table 1), which were then applied to the HBV model in order to maximize Kling Gupta Efficiency (Gupta et al., 2009) of simulated daily discharge. After the GAP has finished, the optimized parameter set were fine-tuned using Powell's quadratic convergent method (Press, 1996) with 1000 additional runs. Lastly, the optimized parameter set was manually adjusted to improve the fits between observed and simulated annual peak flow (see Lamb, 1999). More elaborate calibration and uncertainty estimation procedures such as Generalized Likelihood Uncertainty Estimation (GLUE; Beven and Binley, 1992; Beven, 1993; Beven and Binley, 2014) could be used, but are outside the scope of our study.

After calibration, HBV (two different parameter sets) was used to perform CS with historical CPC and Stage IV rainfall and temperature data to derive long-term simulated soil moisture and snowpack values, which are usually difficult to obtain via measurement. We "pair" samples of these initial conditions with synthetic rainfall events, as described in Sect. 4.2 and Sect. 4.3.

Table 1. Overview of HBV model parameters and prior parameter boundaries.

Parameter	Description	Units	Min value	Max value
Snow Routine				
TT	Threshold temperature for liquid and solid precipitation	℃	-3	3
CFMAX	Degree-day factor	mm $d^{-1}℃^{-1}$	0.5	4
SFCF	Snowfall correction factor	-	0.5	1.2
CFR	Refreezing coefficient	-	0.01	0.1
CWH	Water holding capacity of the snow storage	-	0.1	0.3
Soil Moisture Routine				
FC	Maximum soil moisture storage (field capacity)	mm	100	550
LP	Relative soil water storage below which AET is reduced linearly	-	0.3	1
BETA	Exponential factor for runoff generation	-	1	5
Response Routine				
PERC	Maximum percolation from upper to lower groundwater box	mm d^{-1}	0	10
UZL	Threshold of upper groundwater box	mm	0	50
K0	Recession coefficient 0	d^{-1}	0.5	0.9
K1	Recession coefficient 1	d^{-1}	0.15	0.5
K2	Recession coefficient 2	d^{-1}	0.01	0.15
Routing Routine				
MAXBAS	Length of triangular weighting function	d	1	2.5

We have also added “Section 5.2 Model Validation” by modifying the original paper, P13-14, to:

5.2 Model Validation

We validated the performance of HBV continuous simulation with respect to flood seasonality, frequency of annual daily discharge maxima, and normalized peak flow (i.e. the simulated or observed daily discharge divided by the 2-year flood), using both Stage IV and CPC as precipitation inputs (Fig. 4). We also validated two structures: one with and the other without the HBV snowpack module. The purpose for this latter validation effort is to highlight the importance of proper process representation (and subsequent validation) in process-based FFA.

Simulated flood seasonality varies substantially during the CPC period of record (1948-2016) depending on the inclusion of the snowpack routine. Differences are less for the Stage IV period of record (2002-2016), due to the decreasing role of snowpack in deriving the floods in recent years (Fig. 4a). In both cases, the seasonality of flooding simulated using HBV is improved with the inclusion of the snowpack module, with a higher (lower) frequency of springtime (summertime) floods which more closely resembles observations. Empirical (i.e. plotting position-based) distributions for the simulated annual daily discharge maxima are mostly within the 90% confidence interval (obtained by nonparametric bootstrap) of the observations (Fig.

4b). The CPC-based simulations differ considerably depending on the inclusion of the snowpack module for more common events, but differences in simulated maxima vanish as flood magnitude increases (e.g. AEP<0.1). This is because the most extreme flood events occur later in the season and are thus independent of snowpack or snowmelt processes. Differences are generally negligible between Stage IV-based simulations with and without snowpack, since floods in this shorter, more recent period are generally driven by summertime thunderstorms. These findings are consistent with the general understanding of the regional seasonality of flooding in the region, as discussed in Sect. 5.1.

We compared all simulated and observed flood peaks that can be associated with a USGS observed daily streamflow value that is at least three times the mean annual daily discharge (Fig. 4c). When associating simulated and observed flood peaks, we look within a 2-day window to allow for modest errors in simulated flood peak timing. All peaks in Fig. 4c are normalized by the median annual (i.e. 2-year) flood, which, as a rule of thumb, can be considered as the "within bank" threshold. Again, HBV with the snowpack routine outperforms the model without it, especially for the small to modest flood events in CPC-based simulations. The model without snowpack routine underestimate the small to modest flood events in two cases due to the neglect of water flux from potential snowmelt. While modest scatter exists in the Stage IV-based simulated peaks, there is no obvious systematic bias with event magnitude when the snowmelt routine is included.

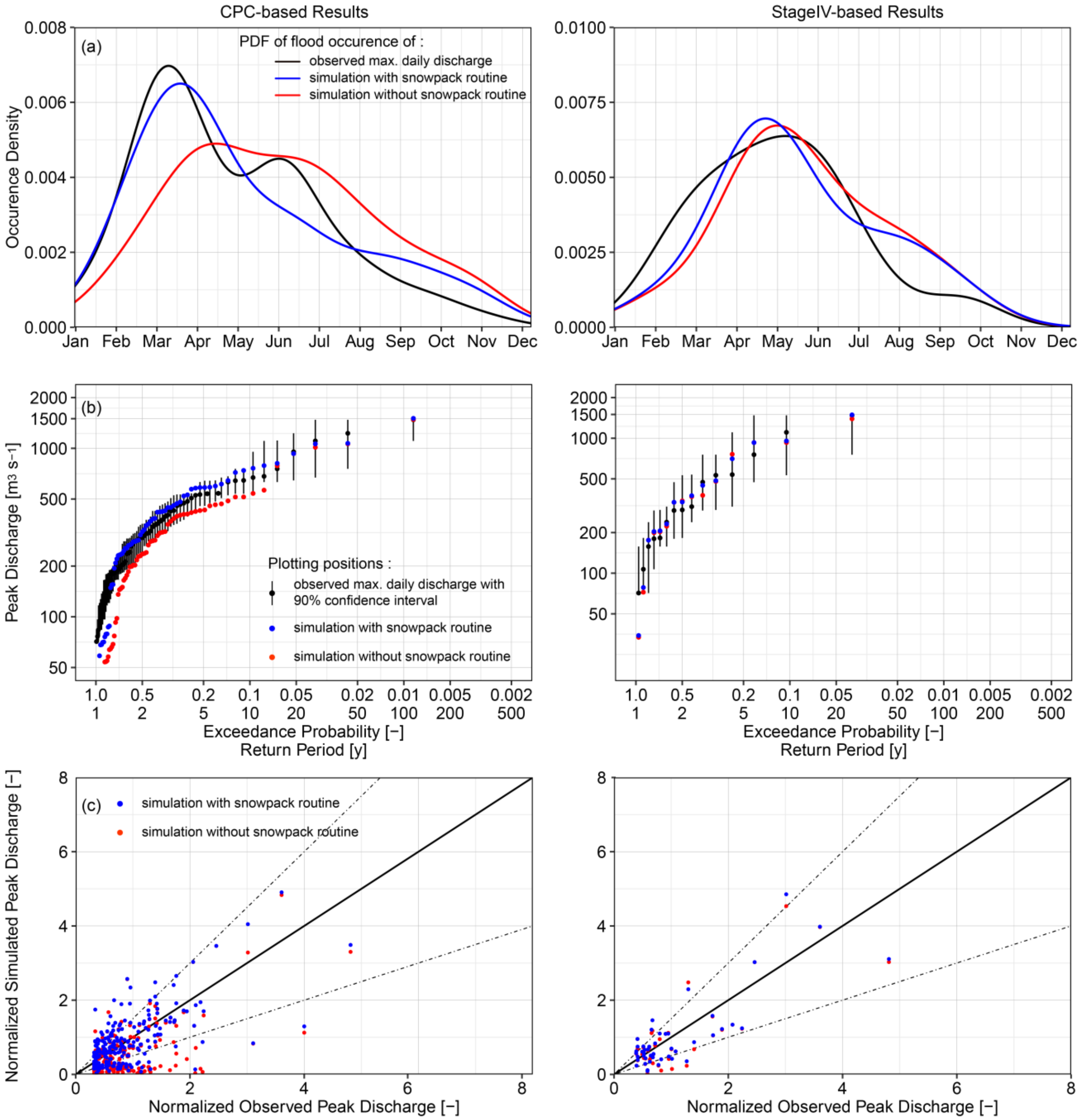


Figure 1. HBV model validation for flood seasonality **(a)**, frequency of annual max. daily discharge **(b)** and normalized peak flow (c). For each panel, the corresponding model validation is performed against CPC- (1948-2016) and StageIV-based (2002-2016) simulation and the results derived from HBV model with (without) snowpack routine are shown in blue (red). The 90% confidence interval for observed max. daily discharge (empirical distribution) is derived using the bootstrapping approach. Peak discharge is defined as a data point with USGS observed value that is at least three times the average observations, and peak discharge are normalized by the median of annual daily discharge maxima (i.e. the 2-year flood). Straight black lines indicate 1:1 correspondence, while dashed lines indicate the envelope within which the modeled values are within 50% of observed.

We also validate HBV's snowpack routine using observed GHCN daily snow depth for two simulation periods (Fig. 5a, 5b) and using USGS daily streamflow observations for Stage IV-based period (Fig. 5c). Because of their differing spatial resolutions and physical representations, point-scale GHCN daily snow depths cannot be directly or quantitatively compared to the watershed-scale snow water equivalent simulated by HBV. Therefore, we validate the snowpack simulation in terms of the snowpack occurrence, defined as the number of occurrences where snow is present on a particular date divided by the total number of years in the historical record. For example, there are 50 days where snowpack is present on January 1st in the 69-year period from 1948-2016, based on GHCN observations and thus the corresponding occurrence rate is 0.72 (50 divided by 69). The HBV model with the snowpack routine captures the central tendency of observed snowpack dynamics, showing that snowpack frequently exists from early November to mid-February, with frequency of snow decreasing from late February until disappearing in early April.

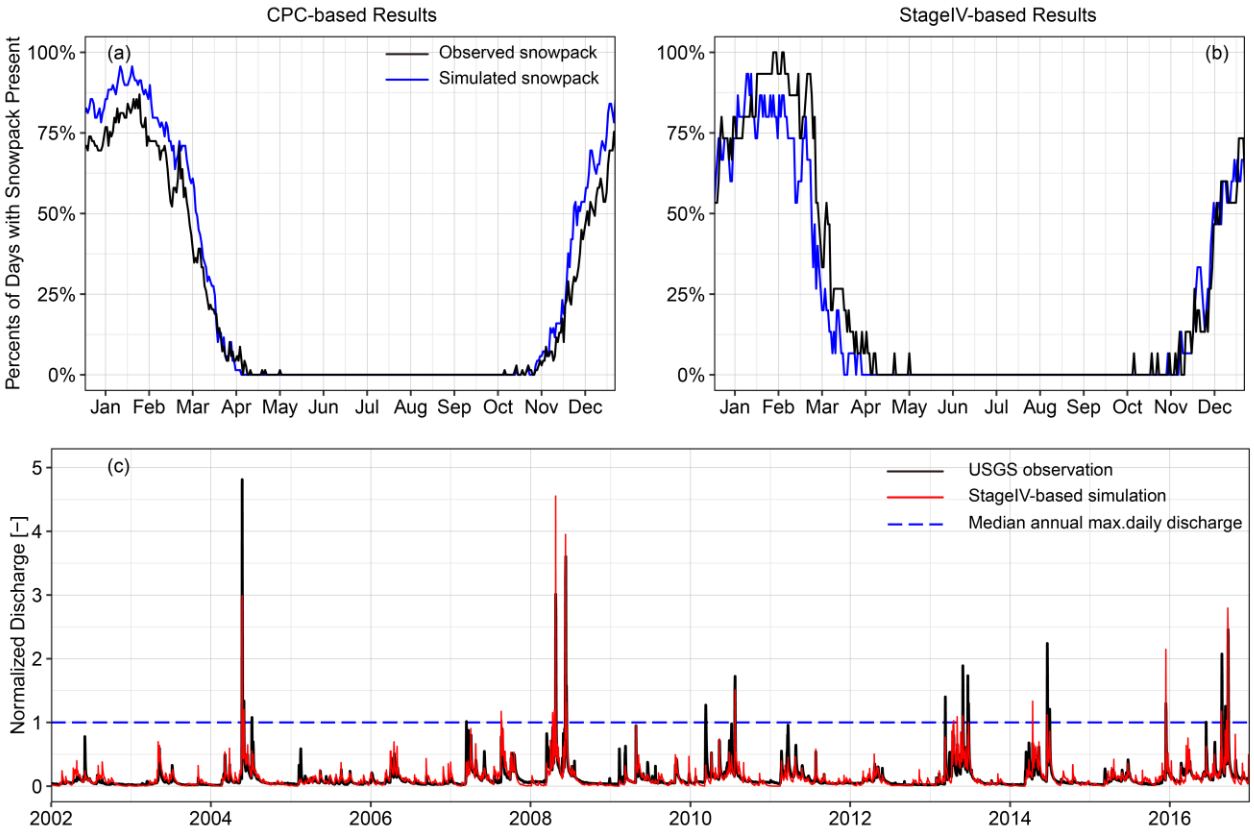


Figure 2. The comparison of percent of days with snowpack present between observations and simulations (a, b) and hydrograph validation for StageIV-based simulation (c). For each day within a year, the percent of snowpack existing days is calculated as the ratio of the number of years when snowpack is present to the total years (69 years for CPC and 15 years for StageIV). Observed and simulated hydrograph are normalized by the median annual flood, which is indicated by the dashed blue line.

Model hydrograph validation is provided in Fig. 5c for the Stage IV period (2002-2016), when major flooding occurred throughout Iowa. Model performance shows no obvious evidence of systematic bias in the streamflow simulations. Although flood seasonality derived by Stage IV-based simulation differs slightly from observations (Fig. 4b), these mismatches are associated with flood events smaller than the median annual flood (blue dash line in Fig. 5c). Stage IV-based simulations do not show bias flood magnitude in late summer. In other words, remaining biases in terms of flood seasonality generally correspond with frequent, small-magnitude events that are typically of less interest in FFA. We therefore conclude that the HBV model with snowpack is generally suitable for subsequent process-based FFA.

The authors explore the utility of hydrological simulations driven by stochastically transposed rainfall fields in deriving flood frequency over a watershed that experiences nonstationarities. Their results highlight the importance of considering changing flood seasonality in flood frequency analysis. While process-based approaches have a fair amount of advantages, their shortcomings are also quite obvious, for instance, mode uncertainty in both parameters and model structure, representation of synthetic rainfall scenarios, etc. As a hydrologist, I would still favor statistical approaches if the gauging record is good (as is the case in this paper). This being said, I would suggest the authors focus on explaining the importance of changing flood seasonality in flood frequency, but rather demonstrating the superiority of process-based approaches to other FFA methods (which is not, as far as I can see).

We thank the reviewer for these useful critiques, that have been very helpful in improving the paper. We fully agree that, particularly in situations of plentiful stream gage observations, statistical approaches are generally preferable. It was never our intention to suggest that our approach is superior to such methods. We note in the original manuscript (P4, line 23-26), however, that there have been prior studies that have demonstrated situations in which rainfall-runoff modeling approaches of various kinds can outperform statistical methods. This, combined with the relative immaturity of rainfall-runoff model-based FFA approaches compared with statistical methods, suggests that additional research, of the kind we present here, can and should be done. As the reviewer stresses, one of the things that such research can point to is the importance of processes and their changes (e.g. seasonal to interannual). In our revised manuscript, we have attempted to emphasize our viewpoint on these issues more clearly. Example revisions to this effect include:

- Include the snowpack routine in the HBV model for both CPC- and StageIV-based simulations.
- Modify the model calibration part (Chapter 4.1) in the original manuscript.
- Add a new section for model validation.
- Address the importance of changing flood seasonality in flood frequency.

We have analyzed two sets of CPC-based results, one for 1948-2016 and the other for 2002-2016 to demonstrate how the changes in flood agents affect the FFA results. We have added the following part to Sect.5.3, P17, line 21 of the original manuscript.

To demonstrate that the discrepancies between the process-based FFA results generated using CPC and using StageIV are driven by changes in flood agents, rather than by differences in model structure (i.e. parameter values), we compared FFA results generated using CPC-based for 1948-2016 and 2002-2016, in terms of event rainfall, initial soil moisture, flood type and peak magnitude (Fig. 8). From 2002-2016 (Fig. 8b), there are fewer flood events driven by snowmelt or rain-on-snow but more driven by rainfall, particularly large magnitude flood events (over 1000 m3/s). In addition, some of the rainfall driven floods (upper left of Fig. 8b) from 2002-2016 indicates high initial soil moisture, which are in accordance with the significant increasing trend of annual precipitation (Table 2). In general, changes in individual flood agents and their interactions can affect flood frequency. Process-based approaches can help illuminate these changes.

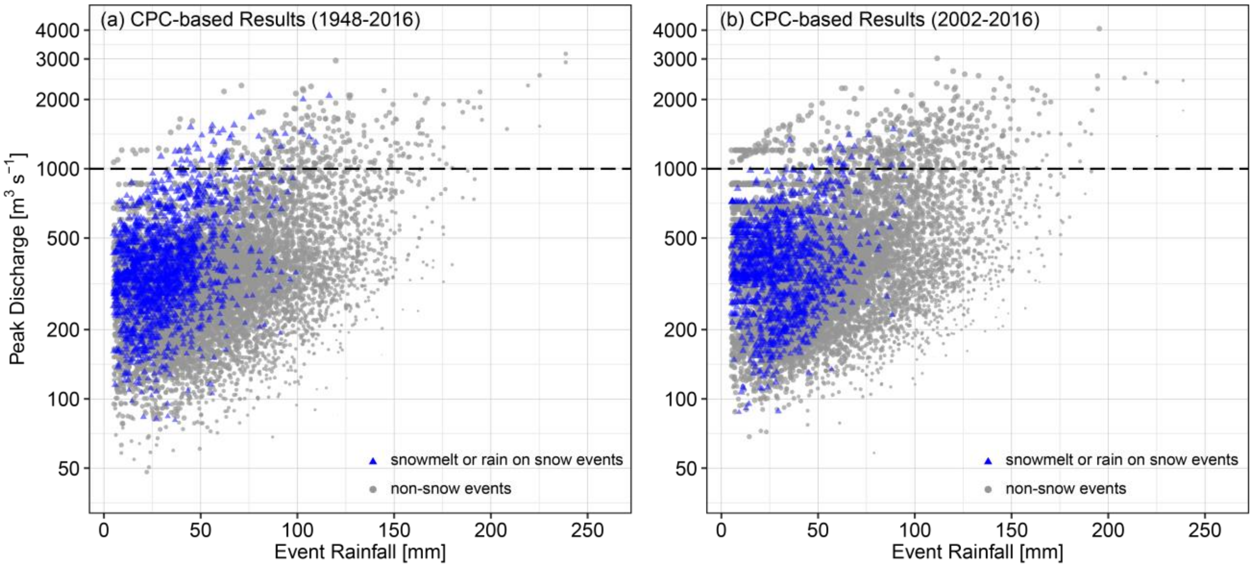


Figure 8. The simulated flood magnitude using CPC rainfall during 1948-2016 (a) and 2002-2016 (b) period, and corresponding antecedent conditions sampled from the continuous simulation. The blue triangles represent the snow related flood events (e.g. snowmelt or rain on snow) and grey dots represents the non-snow related flood events (e.g. rainfall driven). The size of the triangles or dots indicate the antecedent soil moisture with higher value in larger shape. The black dash line indicates the 1000m3/s flood magnitudes.

Specific comments 1-1: An important part is missing from the present paper is model validation. Evidence needs to be explicitly presented to show the capability of long-term model simulations in capturing, for instance, flood seasonality, as well as other features (distribution of annual maximum discharge). This can be done by adding simulation results into Figure 3b and Figure 5a.

We thank the reviewer for this suggestion, which was also voiced by the other reviewers. We have include the model validation, as shown at the beginning of this response, to further demonstrate the capability of long-term simulation in capturing the flood seasonality, high flow magnitude and distribution of annual maximum discharge. We hope the reviewers find it to be more convincing that the limited validation that we included in the original manuscript.

Specific comments 1-2: The authors show a larger frequency of floods during post-summer season in their simulations, could this be possibly related to the positive model biases in representing rainfall-runoff processes during this season? The reliability of process-based approaches in FFA builds on decent model simulations. The authors should spend additional efforts in demonstrating this in the paper. This can be done by providing a quantitative assessment of the model performance.

The hydrograph validation plot (Figure. 5b), along with the flood seasonality validation plot (Figure. 4a) shows that the HBV model with snowpack routine can capture the observed flood seasonality and daily streamflow in the long-term simulation. Although model simulates more flood events in late summer (August-September), it is not biased in terms of late summer flood magnitude. Therefore, we believe these simulated extreme late summer flood events (over 1500 m^3/s) are associated with the regional late-summer storm events in Iowa, rather than model bias.

Specific comments 1-3: Another question about the simulation, how is channel flow represented/considered in the analyses. Antecedent streamflow in the channels can be an important element in representing antecedent watershed wetness, in addition to soil moisture, that plays a role in streamflow simulation.

The reviewer is correct in general that this should be considered within our framework. The HBV model, however, does not need to sample channel flow (streamflow) for the antecedent conditions. The following equations show how the HBV model calculates the streamflow.

$$Q[t] = Q_0[t] + Q_1[t] + Q_2[t]$$

$$Q_0[t] = K_0 * MAX(SUZ[t-1] + recharge[t] + excess[t] - UZL, 0)$$

$$Q_1[t] = K_1 * (SUZ[t-1] + recharge[t] + excess[t] - UZL - PERC)$$

$$Q_2[t] = K_2 * (SLZ[t-1] + PERC)$$

Where conceptually,

$Q[t]$ is the current time streamflow

$Q_0[t], Q_1[t], Q_2[t]$ are the current time overland flow, intermediate flow and baseflow

$recharge[t], excess[t]$ are the current time flux to groundwater and excess runoff, all of which depend on the soil moisture at previous time step

$SUZ[t-1], SLZ[t-1]$ are the water level in upper and lower groundwater box at the previous time step

$K_{0,}K_1, K_2, UZL, PERC\ are\ model\ parameters$

In general, the current time overland flow ($Q_0[t]$) and intermediate flow ($Q_1[t]$) only depend on soil moisture and water level in the upper groundwater box at the previous time step while the current time baseflow $Q_2[t]$ depends on the water level in the lower groundwater box at the previous time step. The more details on HBV model structure can be found in the HBV references in original manuscript, P3, line 4-7.

Specific comments 2-1: The representation of synthetic rainfall fields is another key in process-based FFA approaches. The authors mentioned that they chose ‘most intense rainfall events’ within a prescribed domain. How exactly do they define “most intense rainfall events”? Please explain.

The RainyDay software selects the most intense rainfall events within the transposition domain, in terms of rainfall accumulation of duration *t* and with the same size, shape, and orientation of the watershed. For example, the principal axis of the Turkey River watershed in this study is oriented roughly northwest-southeast and has an area of 4002 km^2. In this case, the 450 selected storms from the historical rainfall data are those associated with the 450 highest 96-hour rainfall accumulations over an area of 4002 km^2 with the same shape and orientation as the Turkey River watershed.

We have modified P10, line 13-14 to:

These intense storms are in terms of 96-hour rainfall accumulation and have the same size, shape, and orientation of the Turkey River watershed, which is oriented roughly northwest-southeast and with an area of 4002 km^2. In order to avoid overlapping storms, these selected events must be separated by at least 24 hours.

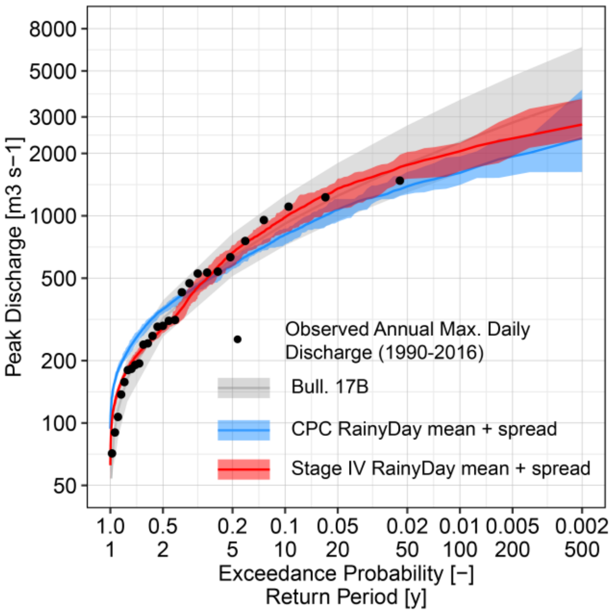
Specific comments 2-2: The authors use the word "realistic" throughout the paper which is inappropriate or miss-leading. They are using synthetic rainfall fields, even though based on real storm events. Please modify.

We believe that our word choice is reasonable when referring the SST-based rainfall fields. They require no parameterization or assumption regarding their spatial or temporal structure (only their starting location is changed), and thus are objectively more realistic than more conventional stochastic rainfall generators. The "realistic" claim would be admittedly more suspect in an environment with complex terrain features (e.g. mountains, coastlines) where both radar estimates and transposition of rainfall fields would be more suspect. Most references on SST in original manuscript, P9, line 30-31, also used word "realistic rainfall".

Specific comments 3: The authors show flood frequency estimates in modern times using Stage IV rainfall fields, and the results match well with gauging records. How about the performance of CPC rainfall in estimating flood frequency?

The RainyDay based FFA using CPC-Unified rainfall data from 2002 to 2016 closely resembles the Stage IV-based FFA, as we mentioned in original manuscript, P16, line 12-15. Regardless, we have added a supplementary plot showing the CPC, Stage IV and Bull.17B based FFA for the modern time (2002-2016).

Supplementary Fig. 1 shows two features that result using CPC data. First, the extreme tail is underestimated, relative to the Stage IV-based simulations and the statistical approach. CPC is known to contain errors in the extreme tail, due to gage undercatch, insufficient gage density to properly sample convective rain cells, and spatial averaging of such cells over large areas, which effectively reduces peak rainfall depths. Second, CPC overestimates the magnitude of more frequent events. This is likely the result of its coarse spatial resolution, which will "smear" rainfall over larger areas (i.e. entire ~600 km2) grid cells when it should be more localized. This would serve to increase the likelihood of rainfall over the watershed, albeit at relatively lower depths/intensities. Thus, if one is to restrict the time period of the rainfall data to recent years (for example, the 2002-2016 time period for which Stage IV is available), then Stage IV would likely be better. As an aside, this belief that Stage IV is preferable to other precipitation datasets in the United States is widely shared in the satellite precipitation community, where Stage IV is often used as a validation dataset.



Supplementary Figure 1. Three peak discharge analyses for Turkey River at Garber, IA: RainyDay with Stage IV (2002-2016) and CPC-(2002-2016) rainfall and USGS frequency analyses (1990-2016) using Bulletin 17B methods. Shaded areas denote the ensemble spread (RainyDay-based results) and the 90% confidence intervals (Bulletin 17B-based analysis), respectively. All observed annual daily streamflow maxima from 1990 to 2016 are shown in black dots.

Specific comments 4: An interesting finding in the paper is described in P17 Line 15-20, but needs to be rephrased. We can see summer floods dominate the upper tail of flood frequency in this region, even though they do not occur as frequent as spring floods. The distribution derived from gauging records is still the 'truth' anyway. Under-representation of summer floods is a pretty common feature of flood peak distributions in the US. I would suggest the authors to provide a brief diagnostic summary of the most extreme flood events in this region.

This is a good suggestion and the newly-added model validation section includes seasonal validations (5.2 Model Validation), as shown at the beginning of this response. Model validations, with respect to flood seasonality, normalized peak flow and hydrograph, show that HBV does not show bias flood magnitude in late summer.

A summary of the most extreme flood events in Iowa are provided in the section 5.1 of the original manuscript, and is provided here: "Flood peak distributions in Iowa "mixtures" of two basic types. Spring floods are associated with springtime rains, high soil moisture, and potentially snowmelt. Summer floods are associated with convective systems. The latter have been shown to significantly affect the upper tail of the flood peak distribution (Villarini et al, 2011) who showed that about 40% of the largest flood peaks are during the May-July period in Iowa. It is important that any process-based FFA approach capture the influence of this mixture on the flood frequency curve." This does not imply that individual gage records are "the truth", only the best representative of it that we have. Thus, discrepancies between model-based approaches and such as ours and observational records warrant further attention.

Specific comments 5: The authors compared simulation results using model with and without snow module, and suggest in the paper that “the modeler must either have sufficient data to diagnose such issues or have sufficient prior knowledge.” (P18 Line 14). I would believe a snow module should be needed in simulation hydrological regimes in this region (dominant spring floods in flood frequency). We cannot simply opt out the snow module by simply checking the simulation. What prior knowledge do the authors have? I would suggest the authors to examine the observed snow climatology over this region, and more ideally, carry out detailed diagnostic analyses of flood agents in this region.

This is a very useful critique. We took this advice into account and developed a new calibration approach that avoids some of the pitfalls that we encountered using more standard calibration techniques. As shown above, we validate this new calibration with respect to flood seasonality, hydrograph, normalized peak flow and snowpack. We finally conclude that the snowpack routine of HBV is indeed important in this study region for this application. We appreciate the insistence of all reviewers in this regard, since it has led to a stronger and more defensible methodology.

Specific comments 6-1: P22 Line5-7, it is not true that conventional statistical FFA methods underestimate flood frequency. At this stage, I would still believe statistical estimates are the ground truth, which enables the evaluation of the process-based approach. The authors do not show updated Bulletin 17B curves using the 1990-2016 flood records in Figure 5, which I would suggest to update.

Figure 5 in the original manuscript shows that conventional FFA methods (defined here as usage of stationary statistical distributions fitted to the period of record using a standard fitting software) underestimate flood frequency beyond the 2-5 year recurrence interval. The statistical fits shown in Figure 5 are included to emphasize that we neglect nonstationarity (as is typically done in FFA practice) at our peril, and usage of “old” data in the face of pronounced hydrologic change can produce incorrect results. We therefore must contend that statistical estimates in such situations should not be considered “ground truth.” Bulletin 17B-based results using 1990-2016 flood peaks are shown in the figure above (see responses to specific comment 3). This fits the observed flood peaks well, as one would expect, though obviously subject to substantial uncertainty for low AEP events due to the short fitting period. Other methods, such as nonstationary FFA, could be used, but our goal is not to prove the superiority of one method or another, but rather to highlight some important issues regarding flood physical processes, their changes, and the resulting implications for flood frequency, issues which are generally ignored in conventional analysis.

As I have mentioned earlier in general comments, it is not wise for the authors to demonstrate the dominating superiority of process-based FFA approaches in this paper, at least for this region. Process-based approach, as presented in this paper (hydrological model + SST), can be highly recommended in poorly gauged watersheds. For poorly-gauged watersheds, however, another issue arises as how to obtain a large ensemble of antecedent watershed wetness conditions used in event-based model simulations. The authors need to provide a discussion about both pros and cons of the proposed approach.

Again, our intention was not to argue for the superiority of process-based methods, and we regret that we gave the reviewer that impression. We have modified the manuscript to make more clear the point that we are attempting to highlight the importance of flood processes and their changes in “shaping” flood frequency, and show an approach that can begin to account for such processes and their changes—though more work is needed, and is ongoing within our research group and elsewhere. Additionally, we agree with the reviewer that a brief discussion about both pros and cons of our framework is necessary.

We have revised the last paragraph of the conclusion to:

A number of issues remain that make broader usage of our process-based framework challenging. Perhaps the biggest limitation of process-based approaches is the necessity of discharge observations, which are central to both identifying hydrologic changes and to calibrate and validate the hydrologic model. Thus, usage of the approach in ungaged basins may not produce satisfactory results. This issue is fundamental to other FFA techniques as well. Statistically-based discharge analyses, for example, similarly rely on streamflow observations, while design storm approaches also require hydrologic model calibration.

Our framework highlights the opportunity and challenge with process-based FFA approaches; namely, that progress on understanding and estimating flood frequency and how it is evolving in an era of unprecedented changes in land use and climate requires better understanding of how the underlying physical processes, and the interactions between them, are changing. Poor model representation of key hydrological processes, however, can lead to incorrect conclusions about present or future flood frequency. Despite the challenge, we share the view of Sivapalan and Samuel (2009) that process-based approaches hold great potential for advances in FFA research and practice, particularly in projecting the future FFA when coupled with high resolution climate model. We do not propose that process-based approaches should necessarily supplant more conventional discharge-based analyses, and discharge observations were central to our present study. Rather, we anticipate a gradual “merging” of statistical and process-based stochastic simulation techniques as well as of the associated observations and synthetic data.

I have a couple of additional comments on word expressions, paragraph organizations, etc., but they can wait till the second round of review. The paper can be a worthwhile contribution to the literature subject to major revisions.

We look forward to further feedback from the reviewer. We have also made minor modification to the structure and word choice in the revised version.