Replies to the comments of Anonymous Referee #3

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	°C	-3	3
CFMAX	Degree-day factor	mm $d^{-1}°C^{-1}$	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.

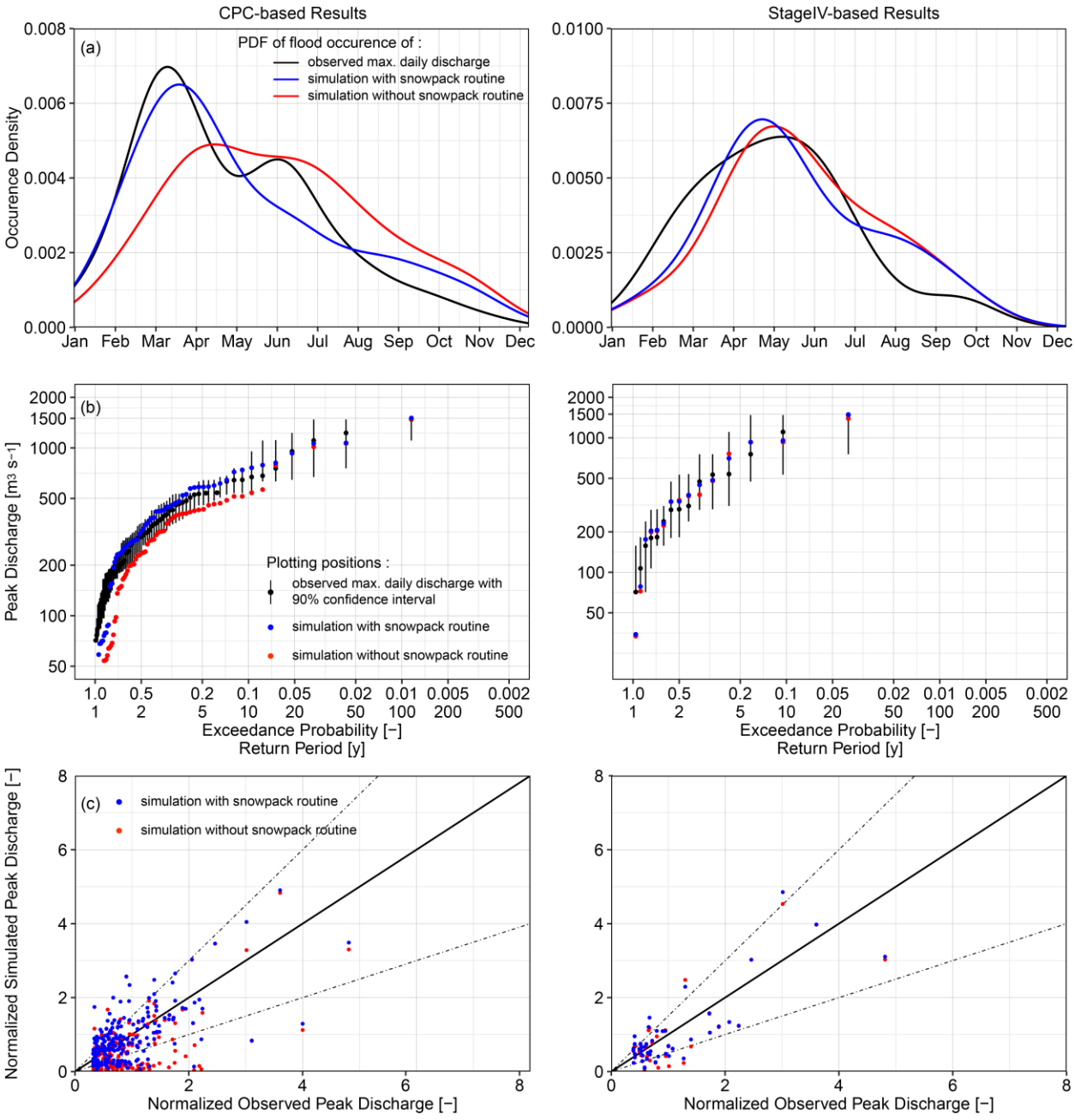


Figure 1. HBV model validation for flood seasonality **(a)**, frequency of annual max. daily discharge **(b)** and normalized peak flow (c). For each panel, the corresponding model validation is performed against CPC- (1948-2016) and StageIV-based (2002-2016) simulation and the results derived from HBV 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.

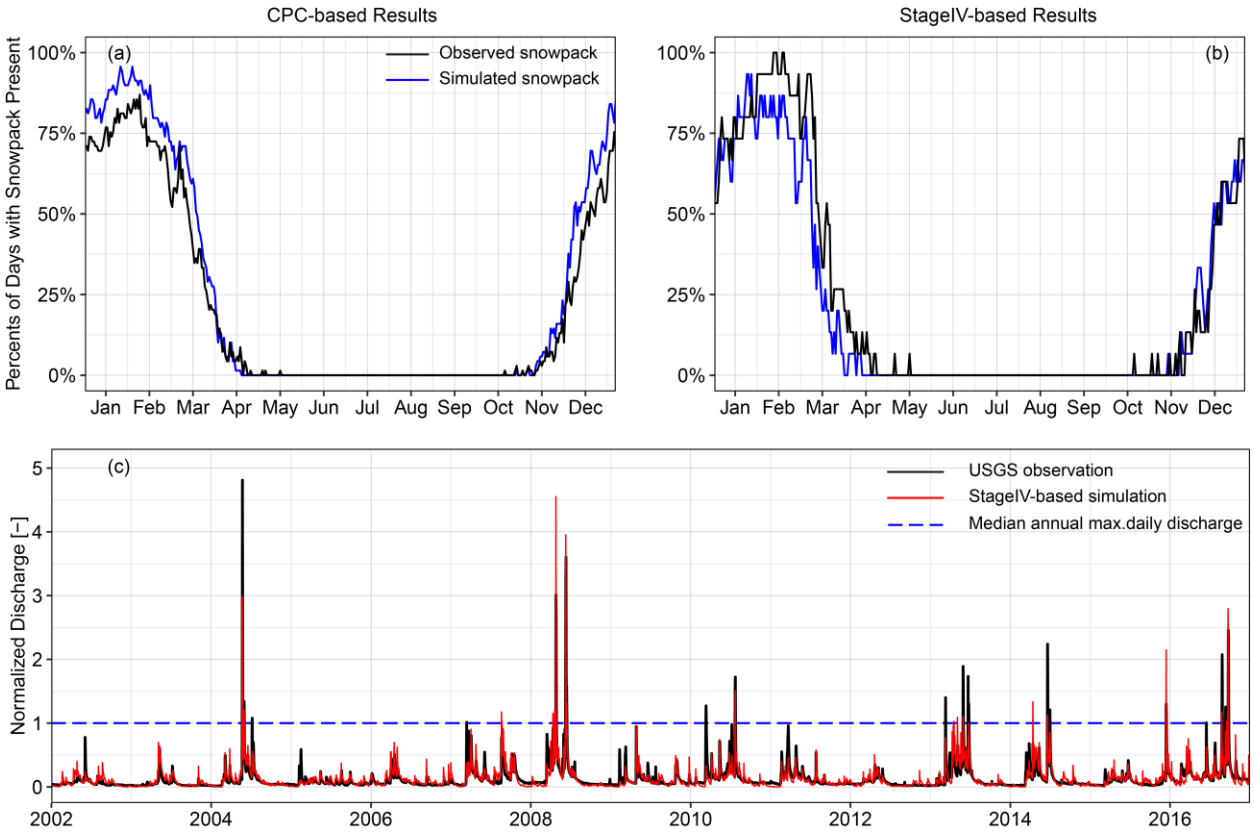


Figure 2. The comparison of percent of days with snowpack present between observations and simulations (a, b) and hydrograph validation for StageIV-based simulation (c). For each day within a year, the percent of snowpack existing days is calculated as the ratio of the number of years when snowpack is 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.

This combination of continuous and event based modelling is a quite novel idea and provides a flexible framework for DFFA. The application of the methods seems sound, the research is done systematically and the paper reads quite well. However, I do have some concerns regarding the selection of the hydrological model, the selection of two precipitation data sets and some of the conclusions. I will detail these below in the major comments, followed by some minor comments. The paper is worth to be published after major revision.

We thank the reviewer for these useful critiques, which have been very helpful in improving the paper. We address these issues more deeply in specific responses, but generally speaking: 1.) in the revised manuscript, we have reintroduced the snowpack routine in the HBV and calibrate and validate the model more carefully. We discussed the model validation with respect to the flood seasonality, peak flow, snowpack, and hydrographs. 2.) we discuss the limitations of CPC precipitation data and the reason why we include the Stage IV precipitation data in this process-based FFA framework. 3.) we provide a short summary of the pros and cons of the proposed FFA framework.

Major comments 1: The selection of the lumped HBV model is not plausible to me, especially given that a) the snow routine is not working and b) the high resolution StageIV rainfall data cannot be utilized by this lumped model.

Since we have updated the HBV model by including the snowpack routine and validated the model as shown in the beginning of this response, we hope the reviewer finds the selection of the lumped HBV model to be more convincing. It also should be noted that, the process-based FFA methodology employed in this study could be coupled with other (sophisticated) hydrologic models, as we mentioned in the original manuscript, P9, line 10, and, in fact, that is our next research direction. Nonetheless, after decades of research, lumped models have still proven to be very useful in a variety of hydrologic fields including flood applications and research. One challenge that we faced in this study was how to quickly implement and evaluate modifications and additions to the methodology, which can be much slower and more challenging using a more sophisticated distributed model.

We respectfully disagree that the Stage IV rainfall data cannot be utilized by a lumped model. Regardless of model choice, Stage IV precipitation data is generally better than CPC data in the study region, in terms of accuracy-this is evident, for example, in the fact that the satellite precipitation community routinely uses Stage IV and related gage-corrected radar products, rather than CPC, to validate satellite rainfall estimates. 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. It is true that the lumped model cannot "leverage" the rainfall spatial structure embedded in Stage IV, but it still benefit from its improved accuracy.

Major comments 2: The application of two rainfall data sets is not plausible and also quite confusing for the reader since a) the Stage IV rainfall data observation period (2002-2016) is covered also by the CPC rainfall data observation period (1948-2016), b) a lumped hydrological model cannot really benefit from high resolution rainfall data (see 1) and c) the hydrological simulation results for both rainfall data sets seem to be very similar (as the authors state on page 16, lines 12-13). I would recommend to do all the simulations with the CPC rainfall if the hydrological model is not changed. If a more suitable hydrological model is selected the two data sets might be kept in the study but the differences in hydrological response using the two data sets for the same time period (2002-2016) need also to be demonstrated and discussed.

We feel that including the Stage IV-based simulation in this case study is important in two respects: 1.) As mentioned in the response to comment 1, we believe the Stage IV precipitation data has high accuracy than CPC. As an aside, this belief that Stage IV is preferable to other datasets when long records are not required is widely shared in the satellite precipitation validation community, where Stage IV is often used as a validation dataset. 2.) We also want to highlight that using only 15 years of rainfall records, our process-based approach can produce accurate estimates of "present-day" flood frequency.

In addition, we have analyzed two CPC-based results from 1948-2016 and 2002-2016 to demonstrate how the changes in flood agents affect the FFAs. 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.

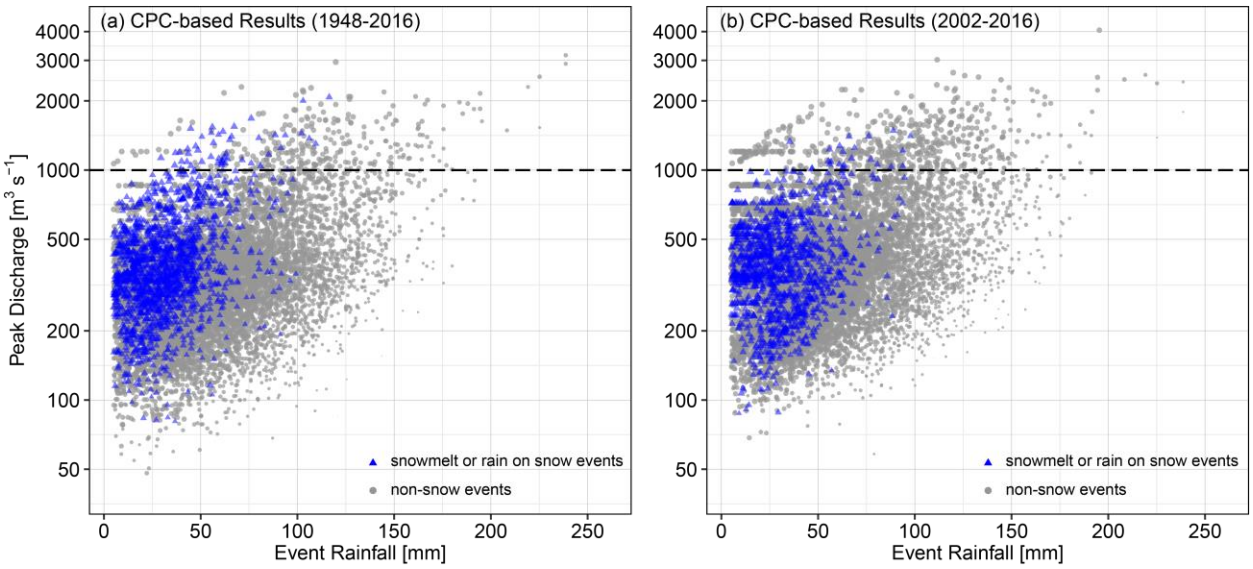


Figure 8. The simulated flood magnitude using CPC rainfall during 1948-2016 (a) and 2002-2016 (b) period, and corresponding antecedent conditions sampled from the continuous simulation. The blue triangles represent the snow related flood events (e.g. snowmelt or rain on snow) and grey dots represents 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.

Major comments 3: The application of a model without snow routine for a catchment with significant snow processes doesn't make sense to me. This way the advantage of process based flood frequency analysis (FFA) is partly lost; obtaining the correct hydrological response for the wrong reason is not satisfying. I am not convinced that the non-stationarity in seasonality is only due to changed soil moisture conditions from rainfall. Temporarily shifted snow dynamics might play a role as well

After taking the reviewers' comments into account very seriously, we recalibrate our model with snowpack routine "turned on" and validate it 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.

Major comments 4: I would be careful with the conclusion, that only with this DFFA method nonstationarity in seasonality can be handled well. Also, non-stationary seasonal FFA approaches are available employing mixed distributions for getting final design values. This needs to be briefly discussed.

We appreciate the comment. Certainly seasonality could be considered using other approaches, though mixture distribution approaches may still not elucidate the fundamental drivers that "shape" flood frequency, even if they can provide good end results. We are not aware of such approaches being used in widespread practice, at least in the United States. Nonetheless, we had added a brief comment in this regard to the conclusions in acknowledgement of this criticism.

We have revised the first paragraph of Section 6 on P21, line 13-15, to :

It must be noticed that the statistical approaches coupling with flood seasonality indices can also investigate the impacts of seasonality on FFA and improve the flood frequency estimation in a regional scale (Ouarda et al., 2006). Our aim is to estimate flood quantiles by reconstructing meteorological and hydrological processes and their interactions, providing an alternative approach which is also well-suited to nonstationary environments (see also Sivapalan and Samuel, 2009).

Major comments 5: This combination of continuous and event based modelling is a good idea. However, there is an important limitation which should at least be mentioned. The framework provides only one possible realization of initial conditions. Nature is more variable. Stochastic rainfall models producing continuous rainfall don't pose this limitation on hydrology.

Each event-based simulation is randomly paired with initial conditions drawn from a continuous simulation (15 years in the case of Stage IV, 69 years for CPC). Thus, we would argue that a large number of possible realizations of initial conditions are used. We would direct the reviewer to Section 4.3. If the reviewer finds this description incomplete, we would appreciate suggestions for how we can make this point more clear. Though we have not tested rigorously, we would guess that relatively short records (say, 15 years) of continuous simulations are sufficient to obtain enough variability in initial conditions. Compared with rainfall, soil moisture (which is bounded between 0 and saturation) and springtime snowpack have thinner tails and thus easier to represent in our framework by sampling from relatively short continuous simulation.

We agree that continuous stochastic rainfall models also have the ability to produce a wide range of pre-event conditions, though it is likely nontrivial to properly calibrate their seasonality with respect to the extreme tail of precipitation-demanding long training datasets.

Minor comments 1: Page 2, line 4: This sentence is confusing. I am assuming you mean '... statistical analysis of observed streamflow, design storms !and! continuous simulation !or! other so called "derived" or "process based" methods'.

Correct. We have modified this sentence to:

Most existing FFA methods belong to one of three approaches: statistical analysis of streamflow observations, design storms, and continuous simulation or other so-called "derived" or "process-based" methods.

Minor comments 2: Page 4, lines 15-17: This sentence seems not to be complete.

We apologize for this. We have revised this sentence to:

Wright et al. (2014a) discusses additional design storm shortcomings including time of concentration concepts, in greater detail, while also pointing out that design storm approaches (like other hydrological model-based FFA) can incorporate future projections in land use and rainfall more explicitly than can statistical discharge-based methods.

Minor comments 3: Page 10, steps 3 and 4: I would stress that the 30 storms per year are randomly transposed over the domain, only sometimes hitting the catchment and sometimes not. They are not all transposed on the catchment, which would lead to an overestimation of the flood frequency. The reader not familiar with your method might misunderstand that.

The reviewer is correct. We have added this sentence to P10, line 22.

It must be noted that some of the *k* transposed storms may not "hit" Turkey River watershed, and thus their calculated watershed rainfall are zero.

Minor comments 4: Page 11, lines 8-9: The selection of the largest event per year for FFA might also be misunderstood. Here, it also needs to be considered that many of the 30 events do not produce any flood if they do not hit the catchment (see comment 3).

We hope the response to previous comments also addresses this one.

Minor comments 5: Page 14: line 2: Should it not be "… but overestimates for pe<0.3 …"

We assume the reviewer mean Page 16, line 2. We have revised this sentence to:

The Stage IV-based flood frequency curve agrees reasonably well with the discharge-based FFA for $p_e > 0.3$ (left panel of Fig. 6), but yields higher estimates for rarer events.

Minor comments 6: Fig. 5: Why did you select the period 1990 – 2016 and not 1980 or 1970 as starting year? This needs to be justified.

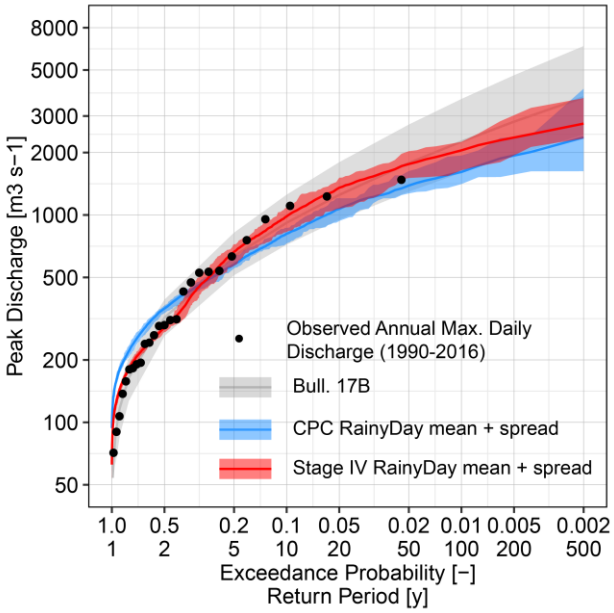
We have not performed any statistical test (e.g. Pettitt test) to determine this change point. However, an "eyeball test" of annual daily discharge maxima (Fig. 1a) from the original manuscript indicates the apparent elevated flood activity during 1990-2016 period. Our arguments do not hinge on a precise determination of when floods in Turkey River began to change, which in any event has likely been a gradual change.

Minor comments 7: Fig. 5: I would also add a statistical analysis (Bull 17.b) for the contemporary period (1990-2016) for comparison.

We have added a supplementary plot showing the CPC, Stage IV and Bull.17B based FFA for the modern time (2002-2016), as other reviewers have suggested.

Supplementary Fig. 1 shows that process-based FFA using CPC precipitation from 2002-2016 closely resembles the Stage IV-based FFA, suggesting that rainfall differences, rather than model structures, are the primary drivers of the differences in this figure. It also 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.



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.

Minor comments 8: Fig. 6: There is no description neither in legend nor in figure caption about the source of the two figures. I assume they stem from different precipitation data sets.

We have updated this figure.