*Replies to the comments of Anonymous Referee #2*

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.

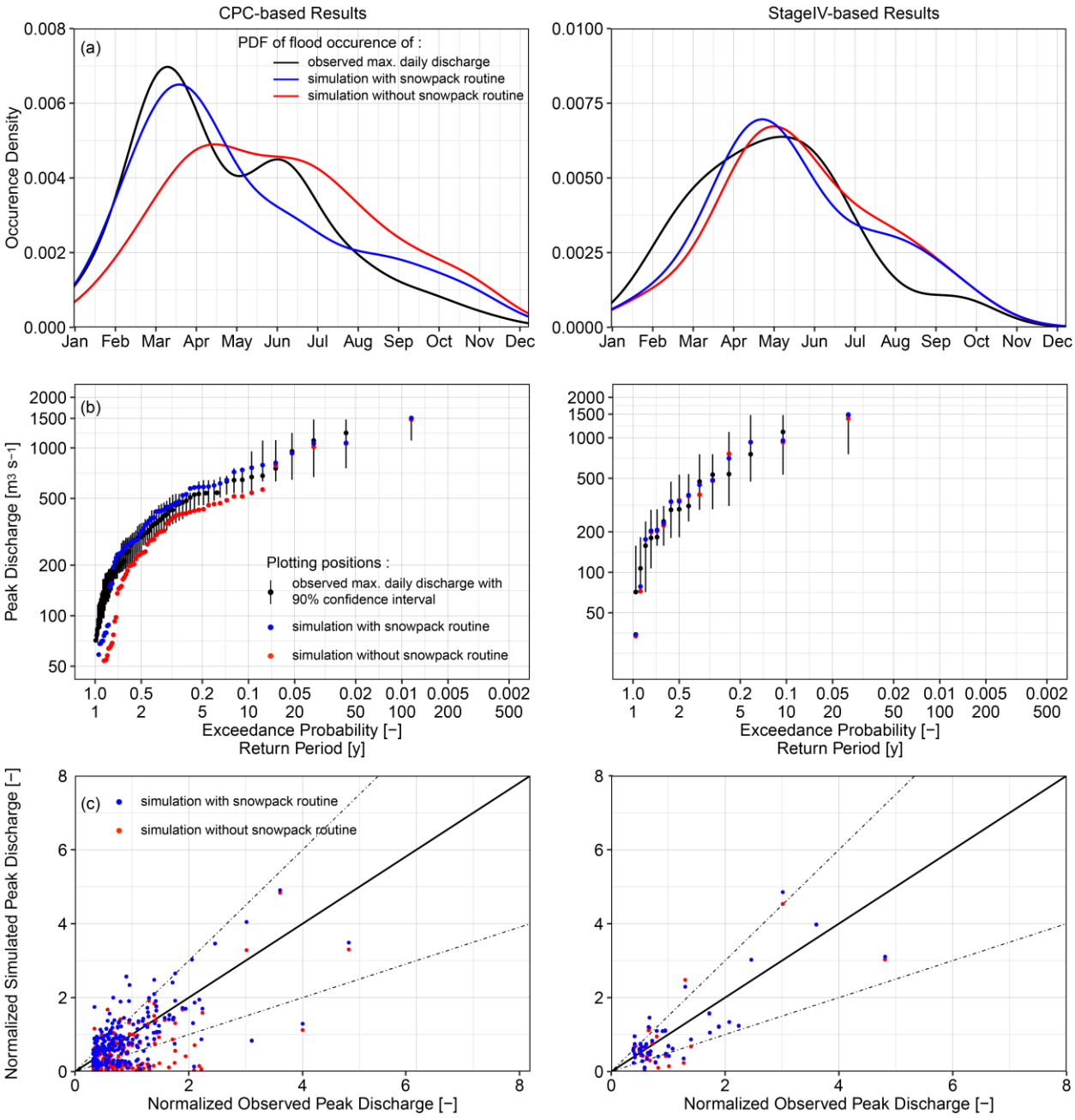



**Figure 1.** HBV model validation for flood seasonality **(a)**, frequency of annual max. daily discharge **(b)** and normalized peak flow (c). For each panel, the corresponding model validation is performed against CPC- (1948-2016) and StageIV-based (2002-2016) simulation and the results derived from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

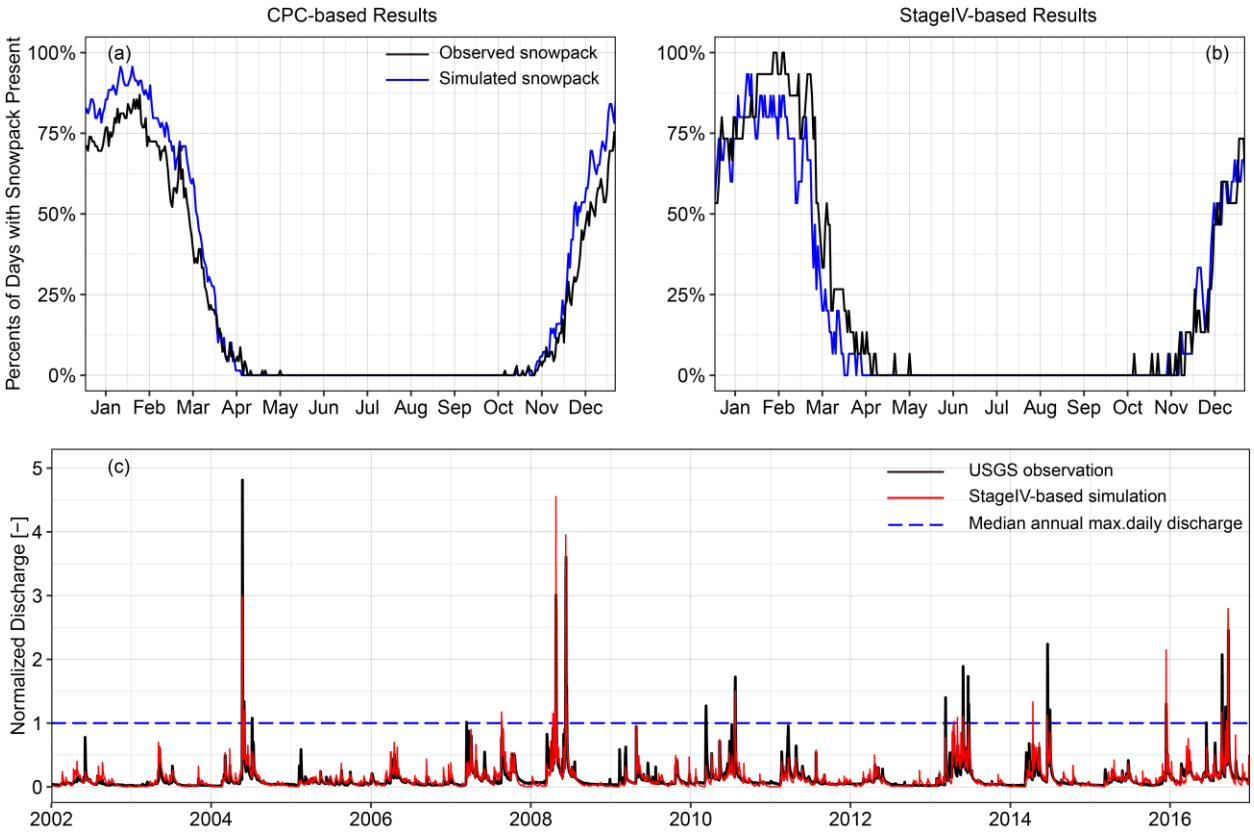



**Figure 2.** The comparison of percent of days with snowpack present between observations and simulations (a, b) and hydrograph validation for StageIV-based simulation (c). For each day within a year, the percent of snowpack existing days is calculated as the ratio of the number of years when snowpack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

The work presents an investigation of flood frequency in the Turkey River basin in the Midwestern United States. The proposed framework, referred to as "process-based" FFA, uses stochastic storm transposition to generate synthetic storms and a lumped hydrologic model to simulate discharge at the outlet of the basin. The authors carry out a series of simulations and corresponding analyses of flood frequency to investigate the impact of seasonality in FFA and potential changes between past and present conditions. Overall, the work has several nice features and the questions posed by the authors are interesting. However, I have some major concerns about certain elements of the proposed framework that need to be addressed before the work can be considered for publication. I provide below major and minor comments that will hopefully help.

We thank the reviewer for these useful critiques, which have been very helpful in improving the manuscript.

Major comments 1: My first and most important concern about the proposed work is related to the choice of the hydrologic model used. The authors mention in different sections themselves that using a lumped model has several limitations. It is good that they acknowledge this limitation themselves but this does not solve the problem. In fact, based on statements as in Line 13, Page 15 "We did not use the snowpack routine…it was shown to produce unrealistic streamflow results" and given that snow processes are important in the selected basins, one immediately recognizes that the choice of the model is not appropriate. If we combine this with the author's statement in conclusions "L22-23, page 22: Poor model representation of key hydrological processes, however, can lead to incorrect conclusions about present and future flood frequency"…I am very skeptical about the conclusions derived based on this model's results. If the model cannot represent well snow processes (particularly flooding due to rain on snow, which should be important in the area) then I fear that the "process-based" FFA is flawed. In this case, the work should be presented at most as a sensitivity analysis and statements such as L1, P22 "helps shed light on the physical processes that shape flood frequency" should be rephrased accordingly.

This is a valid criticism and we thank the reviewer. We hope that the added model calibration and validation, as shown in the beginning of this response, addresses most of the reviewer's present concern. As shown, we have devised a new calibration approach that provided acceptable performance while included the snowpack routine in the HBV model, since we agree with the reviewer that snow processes are potentially important elements of flooding in the region and should not be omitted.

Major comments 2: The calibration and validation of the model lacks clarity. Which forcing was used to calibrate the model? And how the model was validated? These points are not clear in section 4.1. Then in section 5.2 L13,P15 "Different HBV parameters are used…" suggests that separate parameterization was used for the different precipitation forcing but no evidence is provided on a) the validation of the model for the two dataset and b) the variability in model

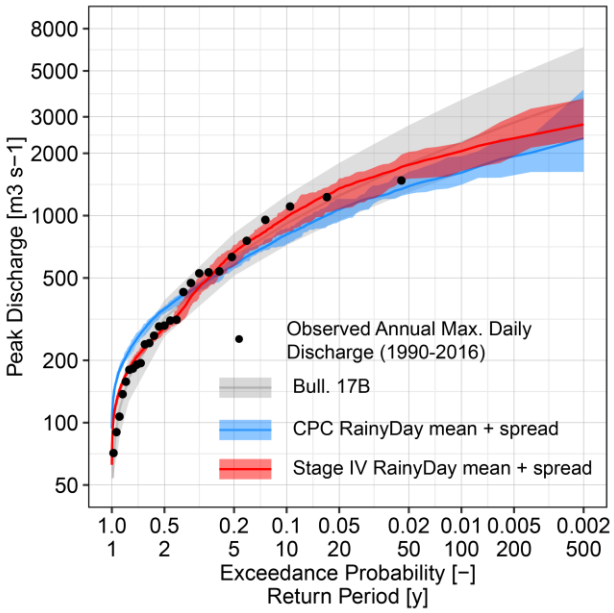
parameters. For the later, if the parameters are significantly different, it will highlight further problems with the approach since this will mean that CPC HBV and CPC-Stage IV simulations treat hydrological processes differently (i.e. may give more weight to different processes in each case). This needs to be investigated and clearly explained in order to understand whether the results can be considered “realistic” or are results of a numerical exercise that mixes two different things.

We hope the updated model calibration can help reviewers find our process-based FFA to be less speculative and more convincing. While ideally model parameters could remain constant regardless of the rainfall dataset used, this is generally not good modeling practice, since rainfall error structures can differ substantially between datasets. For example, due to its much coarser spatial resolution, CPC, even when used in a lumped model, will produce more frequent light rain and lower extremes than Stage IV. Therefore, we believe that calibration for individual input datasets is a necessary evil. Our future research will use distributed physics-based models in place of HBV, and hopefully this is less of an issue in such models.

Major comments 3: For the results in Fig. 5 right panel: Do you use soil moisture years prior to 1990 for the StageIV process-based approach? Also, you should apply the Bull. 17B for the two periods (1933-1989 and 1990-2016) and add them on the graph for comparison.

We did not use the soil moisture prior to 1990 for the Stage IV-based simulation. The antecedent conditions for Stage IV-based simulation are only sampled from continuous simulation of Stage IV period, which is 2002-2016. We have not applied the Bull.17B method to annual daily streamflow maxima for 1933-1989 period because we have not investigated any RainyDay-based simulation for the corresponding time. However, we have added a supplementary plot showing the CPC, Stage IV and Bull.17B based FFA for the modern time (2002-2016), similar to what this reviewer and reviewer 1 suggest.

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 in reality 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 a better choice.




**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 1: P1, L18: “a watershed that is undergoing significant climatic… change”. Is the climatic change at the scale of the watershed only? Consider revising.

We have revised this sentence to:

The methodology is applied to the Turkey River watershed in the Midwestern United States, which is undergoing significant climatic and hydrologic change.

Minor comments 2: P16, L2: “but higher estimates” should be “but gives higher estimates”?

Correct. We have modified that sentence to “but yields higher estimates for rarer events”.

Minor comments 3: Fig.6: Improve caption. What is the upper and what the lower panel?

This figure has been updated.

Minor comments 4: P18L13: “processes in her” should be “processes in his/her”

We have updated the text.