Author's response to the comments of reviewer 2:

Summary: The manuscript employed a large ensemble of hydrologic simulations driven by a climate model (Hydro-SMILE) to investigate the future changes in peak flow characteristics and dynamics. Hydro-SMILE can be a powerful tool for extreme value estimates in hydrology. The approach, based on validated physics-based hydrologic models and large ensemble climate data, is especially helpful for understanding the changes in magnitude, frequency, and dynamics in high-return level peak flow events. However, the authors did not fully utilize the data they generated and the focus of the manuscript has been on how beneficial such a tool can be. Further and more in-depth analyses on the changes in the characteristics of high return level peak flow events are needed to improve the manuscript. See below my major comments. In addition, I provide language edits/suggestions in minor comments for places where I think clarity is lacking.

AC: The authors would like to thank the reviewer for the valuable comments to further improve the manuscript. Detailed author comments (AC) on the reviewer comments (RC) are provided for each individual comment.

Major comments:

1. The manuscript has a title of "…high return levels of peak flows…" but specifically focused on 100-year floods. With a large ensemble, the authors could investigate a range of high return levels of peak flows and see how the frequency, magnitude, and dynamics are projected to change in the future climate.

AC: Thank you for your comment. You are right, the large ensemble would allow for the analysis beyond the 100-year return period, however, despite the extensive data available for analysis, uncertainties calculating return periods beyond e.g., 300-year floods would increase again and the sample size for analyzing the drivers is limited. Further, to be consistent with the storyline of the manuscript and the relevance of the 100-year flood in the study region (i.e., design criteria for flood protection or hydropower infrastructure) we prefer to remain the focus on HF100. We will however extend the analysis of the 100-yr Flood and the climate change impact on the dynamics and flood generating processes, as recommended in your review comment 4. With this extension of the analysis showing results for multiple return periods for all these different topics would lead to complex figures and excessive descriptions which may be repetitive in many places.

2. The manuscript focused on proving that it is beneficial to have a large ensemble to estimate extreme peak flow events (Sections 3.1 and Figures 4 and 5), which, I think, is very obvious so I suggest moving them to the supplementary. The authors have generated a powerful dataset for extreme peak flow estimation, however, there is no analysis of the changes in flood frequency and magnitude. The authors are suggested to substantially expand the analysis on flood frequency and characteristics. See Yu et al. (2020) for an example of flood frequency analysis.

Yu, G., Wright, D. B., & Li, Z. (2020). The upper tail of precipitation in convection-permitting regional climate models and their utility in nonstationary rainfall and flood frequency analysis. *Earth's Future*, 8,

e2020EF001613. https://doi.org/10.1029/2020EF001613

AC: The authors partly agree with the comment on the obviousness of the benefit of large ensemble. We think that this is true for some research areas, such as climate/atmospheric sciences where they can be considered a state-of-the-art tool. However, in hydrology large ensembles are still rarely used and rather statistical methods (e.g., weather generators, emulators) are more frequently used to artificially enhance the sample size. We however think that the use of large ensembles in hydrological studies brings many advantages over these statistical methods, especially when it comes to changes in the dynamics. Thus, we think that highlighting the benefits of the represented approach is still necessary and valuable to make these tools more known and accessible within the hydrological community.

Regarding your comment on "…there is no analysis of changes in flood frequency and magnitude." We disagree and point to figures 6 and 7 depicting the changes in frequency and magnitude (intensity), although presented differently compared to your mentioned reference Yu et al., 2020. Yet, the authors will consider the suggestion to further elaborate on the changes in characteristics (as further mentioned in comment 4) and incorporate additional figures and explanations. For this we further consider shortening the part on the benefits (e.g., move Fig. 5 to the supplement materials).

3. The authors only gave limited information on the evaluation results of the hydrologic model which are essential for building confidence in the following analysis. I suggest including figures and/or tables showing the evaluation results such as time series of flow events with other quantitative metrics such as correlation coefficient, % bias, root mean square error, KGE, NSE, etc. One related question would be how the level of trust (LOT) is calculated.

AC: We will add additional information on the model's performance in the supplement materials. Furthermore, the authors will refer the reader to specific publications for additional information about the model performance, as the mentioned metrics and figures have already been published in several papers. We will likely keep this short in the main manuscript and will include some figures in the supplementary material.

4. The changing dynamics in the future climate (Section 3.2) are very interesting and worth digging into. The authors could dig into it with a mechanistic investigation of possible explanations for why they see such changes in future projections. For example, linking snow water equivalent and rain characteristics to the dynamical changes that are projected at the nivo-pluvial stations.

AC: We agree and will provide further analysis and figures showing the changes in flood driving mechanisms which may explain the depicted changes in flood dynamics for the presented flow regimes.

Minor comments:

Lines 57-59: it is unclear how prediction is a reason for challenges in modeling and predicting high flows by Brunner et al. (2021a).

AC: The Authors will remove prediction from the reasons for challenges in modelling and predicting high flows, as this was wrongfully stated by the authors.

Line 71: change "extraordinary" to "extreme".

AC: The authors will exchange the term as recommended.

The paragraph starting from line 73: Needs substantial rephrasing. 1) Rephrase the sentence "This approach of high spatiotemporal resolution for climate and hydrological modeling is computationally demanding." Do the authors mean "This ensemble-based climate and hydrological modeling approach is computationally demanding because of the high spatio-temporal resolution"?

AC: The authors will rephrase the mentioned paragraph for better clarity.

2) Rephrase sentence "However, considering spatially refined catchment features (e.g., slopes, soil characteristics, land use), precise values due to higher temporal resolution, and the application of a SMILE for hydrological modeling supports an enhanced representation of extreme values within models." Do the authors mean high spatio-temporal resolution of hydro-SMILE is particularly valuable for an enhanced representation of extreme values in models because hydro-SMILE considers spatially-refined catchment features at high temporal resolutions?

AC: The reviewer's interpretation of the sentence is correct; the authors will rephrase it accordingly.

3) Rephrase "Thus, this study focuses on the major Bavarian river basins (upper Danube, Main, Inn) with all their tributaries". It is unclear to me what your study area has to do with the above two statements.

AC: The authors will rephrase this sentence to better explain that a high spatio-temporal resolution is beneficial and necessary for the heterogeneity within the study area.

Line 84: Remove "Therefore".

AC: The authors will remove it.

The paragraph that starts from Line 84: add section numbers throughout the paragraph.

AC: The authors will add section numbers as recommended.

Line 85: Confusing sentence. Remove "…to meet the requirements for the hydrological modeling."

AC: The authors will remove this part of the sentence in Line 85.

Line 86: "…hydro-SMILE along..." should be "…along with…".

AC: The authors will change this phrase according to the suggestion.

Line 95: Remove "As a result"

AC: The authors will remove this phrase.

Line 102: "…(up to 1100 mm precipitation sums in the north, 2500 mm in the south; an average temperature of 10 °C in the north, down to 5 °C (-8 °C on alpine summits)…". Are

the authors referring to annual total precipitation and annual mean temperature? Be clear on that. Also, be sure to mention the data sources for these numbers - are they from Poschlod et al. (2020) as well?

AC: The authors will clarify on the actual meaning of the mentioned precipitation sums and average temperatures and will further add references for these numbers as well.

Line 108: "The major river catchments were divided into a total of 98 smaller sub-catchments based on a common interest in flood protection and a more detailed variation in catchment characteristics, using a selection of gauges (Willkofer et al., 2020)." Rephrase this sentence. It is unclear to me whether the 98 sub-catchments were divided based on the spatial distribution of the 98 selected gauges or whether the gauges were selected because of the division of the 98 sub-catchments.

AC: The authors will rephrase the sentence for a better understanding of why the mentioned 98 catchments were selected for this study.

Line 125: Figure caption of Figure 2: Also introduce what are SDCLIREF and WaSiM in the Figure caption.

AC: The authors will add the explanation for SDCLIREF and WaSiM to the Figure caption.

Line 142: What is "T63"?

AC: T63 is a term describing the original grid resolution of the climate model. As it does not contribute to further understanding of the data, the authors will simply remove this term.

Line 149: "Furthermore, the individual members of the CRCM5-LE are considered independent for the hydrological evaluation period from 1981 to 2099, as the analysis of variations in temperature and precipitation over land and ocean shows (Leduc et al., 2019)." This is the first time the authors mention "hydrological evaluation period". This is a confusing sentence. Aren't the CRCM5-LE individual members independent no matter what time period?

AC: We will rephrase this sentence and likely move this sentence to be clearer. Please also see our response to your comment for Line 214 below. For the period that is considered in this paper (1981-2099) the individual members are indeed fully independent of each other. Only at the start of the simulation (1950-) there is a spin-up period which is needed for the members to achieve the full spread of variability and hence to become independent of each other. This spin-up period is inherent from the chosen initialization scheme (macro and micro). The micro initialization (which are introduced by a very small rounding error) all start from the same ocean conditions (within the five different parent simulations) and it takes the system a few years (up to five years) to develop different ocean states which then makes the members fully independent of each other. This spin-up behavior is shown by Leduc et al (2019) for the respective climate simulations used here. We will rephrase this paragraph to clarify this.

Line 152: "…showing regional and seasonal variations in magnitude over Europe (Leduc et al., 2019)." What variables do the authors mean?

AC: The authors will add the variable (temperature and precipitation).

Line 156: add “match” after “were adjusted to”.

AC: We will rephrase.

Line 157: Change “RCM scale” to “RCM grid”. Did the authors do the interpolation onto the RCM grid? If yes, be clear on what interpolation scheme is used.

AC: The authors will change “RCM scale” to “RCM grid” and we will include the interpolation scheme used to upscale the meteorological reference dataset to the RCM grid. We used a mass-conserving approach.

Line 160: “…3-hourly correction factors for every quantile and month”. Unclear how the 3-hourly correction factor is applied.

AC: We will clarify this. Correction factors were determined for each quantile bin for each month and sub-daily (3h) time step. To preserve the ensemble spread, all members were pooled to obtain the correction factors and these factors were subsequently applied to each ensemble member separately.

Sentence starting on Line 162: I think the authors want to stress that bias correction is inevitable. Rephrase it to “Despite the benefits (increasing reliability of climate change projections of the hydrological impact model, reducing bias in mean annual discharge) and shortcomings (disrupting feedbacks between fluxes, modification of change signals, assumption of a stationary bias) of bias correction are highly debatable (e.g., Teutschbein and Seibert, 2012; Maraun, 2016; Ehret et al., 2012; Dettinger et al., 2004; Chen et al., 2021; Huang et al., 2014), bias correction is often inevitable for climate change impact studies (Gampe et al., 2019).”

AC: Thank you for this suggestion, we will rephrase the sentence accordingly.

Line 168: For such a topographically complex region as described in the “Study area” section, I’m concerned the statistical downscaling between grids that are so different (from 12 km in RCMs to 500 m in hydrologic models) will lose important spatial heterogeneity across the domain. Does the mass-preserving approach address this problem?

AC: The mass-preserving approach ensures that through the downscaling no additional precipitation is added or lost. Meaning, when the downscaled result is upscaled to the RCM scale the mass of the RCM grid is matched. The downscaling basically tries to distribute the coarse RCM information to a higher sub-grid scale. We don’t think that the downscaling will alter the spatial heterogeneity. If so, then rather the bias correction will alter the spatial heterogeneity. As with all interpolation schemes, the obtained spatial result is only one of many possible spatial distributions.

Line 171: “The interpolation result was then applied to the SDCLIREF reference fields (Brunner et al., 2021b)”. Unclear. Do the authors mean the SDCLIREF reference fields are also interpolated using the same method?

AC: No, the SDCLIREF reference fields were generated by combining a multiple-linear regression (considering, elevation, slope, longitude, and latitude) and inverse-distance weighting. The interpolation method is based on the interpolation method used by the German

Weather Service (DWD, Rauthe et al. 2013). We will add the necessary information here to clarify this.

Line 194: “minimizing a weighted combination of performance metrics, including Nash and Sutcliff efficiency (NSE; Nash 195 and Sutcliffe, 1970), Kling-Gupta efficiency (KGE; Gupta et al., 2009), the logarithmic NSE and the ratio of root mean squared error to standard deviation (RSR; Moriasi et al. (2007)) (Eq. (1))”. Introducing the overall metric (OM) equation first and then describe what is in the equation. Then, give a threshold - what OM value is considered “good” or “bad”?

AC: The authors will introduce the overall metric first and then describe it. However, the DDS-SA approach will always try to minimize this overall metric regardless of the threshold (although one could introduce a threshold to reduce the number of iterations once the threshold is reached). Once it reaches 0 the algorithm stops. However, provided that an unlimited number of iterations is applied, the algorithm will go on trying to minimize the metric’s result. The authors did not add an additional threshold which terminates the algorithm once this threshold is reached. Thus, we could provide a threshold where we consider the model to be good or bad, but in terms of evaluation this threshold would not be of much significance.

Line 203: “(NSE: 16; KGE: 5)” Unclear. What does this mean?

AC: These numbers mean, that 16 (5) gauges out of 98 did not perform sufficiently well (lower than 0.5) for NSE (KGE). We will clarify this.

Line 208: “Consequently, the level of trust (LOT) for peak flows of return periods of 5, 10, and 20 years of flood events, introduced in Willkofer et al. (2020) showed a moderate to high confindence for most catchments, with gauges of poor simulated performance yielding a lower LOT with increasing return levels.” Do the authors mean gauges with good performance have higher LOT for peak flows with return periods of 5, 10, and 20 years, whereas gauges with poor performance have lower LOT, especially for peak flows at longer return periods?

AC: The reviewer is correct. The meaning of this phrase is misleading, and the authors will thus rephrase this sentence to clarify the meaning of the LOT in this case.

Line 214: “The entire modeling period is shortened by ten years to account for the time span it takes the RCM to produce fully independent realizations due to the inertia of the ocean model (Leduc et al., 2019).” I have two comments: 1) I saw that the authors partially address my question for Line 149 here. It would be better to rearrange this part and the sentence on Line 149 such that the 10-year spin-up period and the choice of the evaluation time period are more clearly lined up and explained. 2) Rephrase this sentence to “We focus on 1961—2099 as opposed to 1950 – 2099 to account for the time it takes for the RCM to produce fully independent realizations due to the inertia of the ocean model (Leduc et al., 2019).”

AC: The authors will rearrange the paragraphs and will combine. Please also see our comment to line 149 above since they are connected.

Figure 3: What is “HF T,BM” on the far right?

AC: This term describes the benchmark (BM) high flow of return period T which is calculated using the empirical probability (p) and the entire series of 1500 yearly peak flows. The authors will describe this value in the Figure caption.

Line 256 and thereafter: Change “intensity” to magnitude throughout the manuscript.

AC: The authors will change the term throughout the manuscript.

Line 293: “…as indicated by the spread of the blue markers around the black benchmark line.” I think the authors mean “… as indicated by the decreasing spread of the blue markers around the benchmark line with increasing sample size.”

AC: The reviewer is correct. The authors will change this phrase accordingly.

Lines 296-297: What are panels a, c, and e?

AC: This is a clear mistake in referencing the different parts of the Figure. The authors will change the references to Fig. 4a, 4b, and 4c.

Figures 1&6: What do “balanced” and “unbalanced” pluvial mean?

AC: According to Poschlod et al. 2020, the term balanced describes a relatively even flow regime with only little variation in monthly mean flow and only little snowmelt influence during November to March. Poschlod et al. 2020 further describes the unbalanced pluvial regime as having a more pronounced peak in mean monthly discharge from January to March and a more pronounced decline during the second half of the year. The authors will add this information to the manuscript.

References:

Poschlod, B., Willkofer, F., and Ludwig, R.: Impact of Climate Change on the Hydrological Regimes in Bavaria, Water, 12, 1599, doi:10.3390/w12061599, 2020.

Willkofer, F., Wood, R. R., Trentini, F. von, Weismüller, J., Poschlod, B., and Ludwig, R.: A Holistic Modelling Approach for the Estimation of Return Levels of Peak Flows in Bavaria, Water, 12, 2349, doi:10.3390/w12092349, 2020

Brunner, M. I., Swain, D. L., Wood, R. R., Willkofer, F., Done, J. M., Gilleland, E., and Ludwig, R.: An extremeness threshold determines the regional response of floods to changes in rainfall extremes, Commun Earth Environ, 2, doi:10.1038/s43247-021-00248-x, 2021