Response to anonymous referee #2

We thank referee #2 for reading the manuscript carefully and providing thoughtful and constructive comments. Below, their comments (C) are reproduced with our responses (R) following. Manuscript text is shown in serif font, intended changes and additions are underscored.

C2.1 Introduction. I have the feeling that only the most standard methods have been referenced in the introduction (conventional flood frequency analysis, regional flood frequency analysis, use of historical information etc.), but in the literature other approaches linking flood estimation with physical processes are present (see e.g., Basso et al. (2021) and references herein for a systematized description of a mechanistic-stochastic physically-based approach for the estimation of river flows/floods). I would suggest the authors to mention other-than-conventional and widely used approaches in the paper introduction, especially if relevant in the discussion of physically-based models or methods less affected by the time series shortness, as the standard ones usually are.

R2.1 Thank you for pointing this out and bringing up alternative approaches.

As for the introduction in general, we indeed tried to mention the most common approaches relevant in our context, namely with focus on applications in large river basins, and on rare to very rare events. We will clarify this and at the same time add some key references concerning flood estimation in a broader context.

As for the mechanistic-stochastic physically-based approach, we will add a paragraph after L55 to mention potential benefits and limitations in the context of our research, following Basso et al. (2021). The preceding text as well as the subsequent introduction to the continuous simulation approach starting on L56 will be slightly reorganised and extended to make it fit to the new paragraph.

Overall, we will modify L40–67 of the introduction as follows:

Generally speaking, common approaches for flood estimation can be categorised into statistical and deterministic (or hydrological) methods as well as combinations thereof (for an overview and evaluation see e.g., Rogger et al., 2012; Okoli et al., 2019). Statistical approaches are widely used (see e.g., Castellarin et al., 2012; Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall, 2012; England, Jr. et al., 2019; Environment Agency, 2020) and also popular to derive design floods for safety assessments. For this, conventional frequency analysis is performed on observed streamflow records, and then a simple return period conversion factor given by design codes (e.g., Bundesministerium für Land- und Forstwirtschaft, Umwelt und Wasserwirtschaft and Technische Universität Wien, 2009; Bundesamt für Energie, 2018; International Commission on Large Dams, 2018) is applied. In addition, it is possible to augment flood frequency analysis with additional data and evidence (Gutknecht et al., 2006; Merz and Blöschl, 2008) such as historical floods (e.g., Bayliss and Reed, 2001; Neppel et al., 2010; Hall et al., 2014; Benito et al., 2015; Salinas et al., 2016; Wetter, 2017), paleofloods (Benito and Thorndycraft, 2005; Baker, 2008; Baker et al., 2010; Benito and O'Connor, 2013; O'Connor et al., 2014), regional frequency analyses (Hosking and Wallis, 1993, 1997), envelope curves (Castellarin et al., 2005), or by differentiating for flood-generating mechanisms (Fischer, 2018; Barth et al., 2019). Also, floods can be estimated from rainfall information via simple approaches such as the GRADEX method (Guillot and Duband, 1969; Naghettini et al., 1996) or the rational method (Mulvany, 1851). Nevertheless, the comparatively short streamflow records contain a rather heterogeneous and likely unrepresentative sample of floods, and neither of the aforementioned methods is able to cover the whole gamut of possible hydrometeorological patterns and the corresponding responses of the river system.

This issue has even greater relevance in large river basins, where flows from individual tributaries interact in a complex manner (see Guse et al., 2020), possibly further complicated through flow management (e.g., lake regulation and reservoir operation).

While the above approaches are predominantly based on statistical elements, further approaches have emerged that combine random elements with understanding of the most relevant physical factors such as soil moisture and runoff dynamics (see e.g., Laio et al., 2001; Porporato et al., 2004; Botter et al., 2007, 2009; Basso et al., 2015, 2016; Zorzetto et al., 2016). Linked with a systematic description of advances in this field, Basso et al. (2021) recently introduced the PHysicallybased Extreme Value (PHEV) distribution as an example of such a mechanistic-stochastic and physically based approach. PHEV showed lower uncertainty and less bias in estimation of large floods (return period of 1 000 years, daily time scale) in comparison to conventional frequency analysis, albeit with a tendency for a slight underestimation and higher variability in performance. Main limitations of PHEV are the assumption of an invariable recession coefficient as well as the exclusion of some hydroclimatological regimes (in particular snow- and glacier-dominated, monsoon and seasonally dry).

Another common approach used in safety assessments are PMP-PMF (Possible Maximum Precipitation-Possible Maximum Flood) estimates, which can follow deterministic (hydrometeorological) or statistical concepts (World Meteorological Organization, 2009). This approach can achieve the range of peak flow extremes examined here, but results have no clear estimate of return period and are usually not applicable over large spatial domains. Moreover, the estimation of PMP and ensuing PMF bears substantial simplifications and considerable uncertainties (Salas et al., 2014; Micovic et al., 2015; Ben Alaya et al., 2018; Zhang and Singh, 2021)

Hydrological methods avoid the abovementioned limitations, more comprehensively link flood estimation with physical processes, and allow for representing effects caused by operation of hydraulic infrastructure. Such methods typically involve a catchment runoff model that is fed with meteorological data and provides simulated discharge as an output (Beven, 2011). In case continuous simulation (CS) is employed rather than an event-based approach, there is no need to separate discharge into baseflow and stormflow, and assumptions about antecedent conditions of a flood event (e.g., snowpack, soil moisture, storage levels of lakes and reservoirs) are not required (Calver and Lamb, 1995; Pathiraja et al., 2012). Beven (1987) was one of the first to recognise the potential of this compelling approach, and CS has indeed been implemented in numerous studies since. However, application in industry is still challenging due to the considerable effort necessary (see overview by Lamb et al., 2016 and references therein). In CS, precipitation data are required to perform rainfall-runoff simulations and subsequently process the simulation results with conventional frequency analyses. Although observed series of precipitation can be used […]

C2.2 Study area and observational data. Why do time series end in 2014? Are there no more recent data available?

R2.2. The research presented in this paper was done within the EXAR project, for which work started in early 2016. At that time, consolidated data were available until the end of 2014 only.

C2.3 Methods. What is exactly the rationale behind the choice of using two different weather generators? I understand that they are implemented independently, and they are used for different purposes, but I miss a clear explanation of the reasons why for example you choose GWEX instead of the SCAMP as input to the HBV model and not the other way around.

R2.3 The two different weather generators are first of all used to assess the structural uncertainty in the meteorological part of this study. They are actually used for similar purposes but, as indicated on L266–270 in the original manuscript, it was only possible to run the full set of

GWEX generated weather scenarios through HBV light and RS Minerve due to the exceptionally high computational cost of long simulations at multiple sites.

Please see also in particular our response to comment C1.8 by referee #1 (who also addressed the rationale for using two different weather generators): For clarification, we will move Lines 710–713 to the section containing Lines 156–159, slightly adjust the text and add cross-references:

Our model chain consists of three main components. First, two weather generators – GWEX (Section 3.2.1) and SCAMP (Section 3.2.2) – were used to provide 30 time series scenarios of precipitation and temperature with a length of 10 000 years each, and to assess the structural uncertainty in the meteorological part of this study (Sections 5.3 and 5.5). Second, the full outputs of GWEX were used as input for the bucket-type catchment model HBV [...]

C2.4 L193: It is not clear to me what do you mean by "…represents the dependence structure of innovations in the generation process"

R2.4 "Generation process" refers to the multivariate autoregressive model (MAR) process which contains so-called innovations which are modelled using a multivariate distribution. "Generation" will be replaced by "multivariate autoregressive model (MAR)":

A Student copula represents the dependence structure of innovations in the multivariate autoregressive model (MAR) and introduces a tail dependence between at-site extremes.

C2.5 L268: could you be clearer about the "technical issues"? What are they related to?

R2.5 This point was also raised by referee #1, see our response to comment C1.19: The most likely cause of the issue was a file transfer problem, i.e., the silent crash of a copy process of GWEX data to the computational cloud where HBV was run. Due to this, the HBV simulations for the 11 000 scenario years in question were forced with data from an outdated version of GWEX. The ensuing inconsistency was only discovered a few months later, when further computationally intensive work had been done within the EXAR project on the basis of the continuous simulations. It was therefore decided to discard the affected scenario years, as the remaining consistent simulations were still 289 000 years long and thus sufficient for the scope of EXAR. We will complement the text as follows:

From the 300 000 years simulated in total, 11 000 were discarded due to an inconsistency (most likely caused by a file transfer problem and the subsequent usage of an outdated version of GWEX data; for details, see Viviroli and Whealton, 2020), leaving 289 000 years for detailed analysis.

C2.6 Results. There is not a clear discussion about the reasons why the two weather generators provide different precipitation ranges. I think that you should spend some time on better describing the differences in the outputs obtained through the simulations and what they are related to.

R2.6 This is a relevant comment which was also raised by referee #1, please see our response to comment C1.22: We will extend and rephrase the last paragraph of Section 4.1.2 as follows:

At the scale of the entire Aare River basin, MAP extremes are roughly similar for GWEX and SCAMP (Figure 3, Figure 4). At the sub-basin scale, however, the extremes of SCAMP are generally larger than those of GWEX and show slightly different spatial patterns. Both of these differences are probably explained by the fact that the two weather generators are built upon substantially different approaches and generation processes: GWEX produces multi-site 3-day amounts disaggregated at a daily scale,

whereas SCAMP produces regional MAP and MAT values at a daily scale. Three-day maxima in SCAMP are thus the result of the aggregation of three consecutive daily simulated values. The temporal coherency between MAP values generated by SCAMP for consecutive days comes from the large-scale atmospheric forcing, which follows relevant atmospheric trajectories from one day to the next. However, this conditioning does not necessarily preserve the day-to-day dynamics of rainfall systems. Nevertheless, it can be noticed that the largest difference – found for the Neuchâtel sub-region – is rather moderate (+10% for MAP3d and +20% for MAP1d). A further comprehensive evaluation of precipitation time series generated with both weather generators is found in Evin et al. (2018, 2019) and Chardon et al. (2020), as well as in Raynaud et al. (2020), which reports on severity, spatial and temporal dynamics, and meteorological relevance of events.

C2.7 L370: I would avoid reporting only the Nash-Sutcliffe efficiency values, but at least complement them with another evaluation criterion, as the NSE is not the optimal one when model accuracy needs to be assessed.

R2.7 That is of course a valid point, thank you. We will report on all three efficiency criteria shown in Figure 6, but for brevity declare median efficiencies. The range can be inferred from Figure 6. We will in addition report these median efficiencies for the three sites in the Emme, Lorze, and Saane Rivers that showed poorer performance:

Results (Figure 6b) show good to very good agreement between observations and simulations (median efficiencies over all three representative parameter sets: NSE 0.83, KGE 0.85, KGE_NP 0.89) for all sites in the Aare, Reuss and Limmat Rivers. The three sites in the Emme, Lorze, and Saane Rivers showed poorer performance (NSE 0.34, KGE 0.65, KGE_NP 0.66).

The abbreviations of the efficiencies used above are currently only declared and used in the legend for Figure 6. Therefore, we will define them also in the main text on L437ff.:

Hydrological simulations for the individual HBV sub-catchments were evaluated based on three criteria: the Nash-Sutcliffe (NSE) (Nash and Sutcliffe, 1970), the Kling-Gupta (KGE) (Gupta et al., 2009) and the non-parametric Kling-Gupta (KGE_NP) (Pool et al., 2018) efficiencies.

C2.8 L418: I suggest the authors to define the FOEN acronym, as it is not clear what you are referring to (I had to go to the Acknowledgments section to understand its meaning).

R2.8 Thank you for noting this. In response to comment C1.25 by referee #1, we will expand the description of extrapolation methods in Chapter 2 on study area and observational data, and define the FOEN acronym there already.

C2.9 Figure 10a. Despite considering this representation very nice, I have to say that most of the information in the smallest circles is lost. I would suggest leaving it like it is for the entire basin and the sub-regions but simplify the symbols for all the other sites (maybe only showing a couple of representative durations, so that the colors are clear).

R2.9. We understand this comment and recognize that it is difficult to appraise the information in the smallest circles precisely. However, we argue that the regions with the largest return levels still stand out thanks to all the small circles. This information cannot be inferred in similar detail at the level of sub-regions only. We have tried a number of alternative design solutions, but the current representation still appeared to be best for highlighting the multiscale structure of precipitation extremes, as well as the regions and areas with the highest precipitation amounts.

C2.10 L470: I believe the first comma should be removed

R2.10 Thank you, will be done.

C2.11 L666: a comma is missed after e.g.

R2.11 Thank you, will be added

References

Barth, N. A., Villarini, G., and White, K.: Accounting for Mixed Populations in Flood Frequency Analysis: Bulletin 17C Perspective, J. Hydrol. Eng., 24, https://doi.org/10.1061/(ASCE)HE.1943-5584.0001762, 2019.

Basso, S., Schirmer, M., and Botter, G.: On the emergence of heavy-tailed streamflow distributions, Adv. Water Resour., 82, 98–105, https://doi.org/10.1016/j.advwatres.2015.04.013, 2015.

Basso, S., Schirmer, M., and Botter, G.: A physically based analytical model of flood frequency curves, Geophysical re-search letters, 43, 9070–9076, https://doi.org/10.1002/2016GL069915, 2016.

Basso, S., Botter, G., Merz, R. and Miniussi, A. (2021). PHEV! The PHysically-based Extreme Value distribution of river flows. Environ. Res. Lett., 16 (12). doi:10.1088/1748-9326/ac3d59

Beven, K. J.: Rainfall-Runoff Modelling: The Primer, 2nd edition, Wiley, Chichester, 2011.

Botter, G., Porporato, A., Rodriguez-Iturbe, I., and Rinaldo, A.: Basin-scale soil moisture dynamics and the probabilistic characterization of carrier hydrologic flows: Slow, leaching-prone components of the hydrologic response, Water Re-sour. Res., 43, 181, https://doi.org/10.1029/2006WR005043, 2007.

Botter, G., Porporato, A., Rodriguez-Iturbe, I., and Rinaldo, A.: Nonlinear storage-discharge relations and catchment stream-flow regimes, Water Resour. Res., 45, https://doi.org/10.1029/2008WR007658, 2009.

Calver, A. and Lamb, R.: Flood frequency estimation using continuous rainfall-runoff modelling, Physics and Chemistry of the Earth, 20, 479–483, https://doi.org/10.1016/S0079-1946(96)00010-9, 1995.

Castellarin, A., Kohnová, S., Gaál, L., Fleig, A., Salinas, J. L., Toumazis, A., Kjeldsen, T. R., and Macdonald, N.: Review of Applied European Flood Frequency Analysis Methods, COST Action ES0901, WG2, Wallingford, Oxfordshire, UK, 130 pp., 2012.

Deutsche Vereinigung für Wasserwirtschaft, Abwasser und Abfall: Ermittlung von Hochwasserwahrscheinlichkeiten: DWA-Regelwerk, Merkblatt DWA-M, 552, Hennef, 90 pp., 2012.

England, J. F., Jr., Cohn, T. A., Faber, B. A., Stedinger, J. R., Thomas, W. O., Jr., Veilleux, A. G., Kiang, J. E., and Mason, R. R., Jr.: Guidelines for Determining Flood Flow Frequency: Bulletin 17C, Version 1.1, May 2019, U. S. Geological Survey Techniques and Methods, Book 4, Chapter 5b, 168 pp., 2019.

Environment Agency: Flood Estimation Guidelines, Technical guidance, 197 08, 129 pp., 2020.

Fischer, S.: A seasonal mixed-POT model to estimate high flood quantiles from different event types and seasons, Journal of Applied Statistics, 1–17, https://doi.org/10.1080/02664763.2018.1441385, 2018.

Laio, F., Porporato, A., Ridolfi, L., and Rodriguez-Iturbe, I.: Plants in water-controlled ecosystems: active role in hydrologic processes and response to water stress, Adv. Water Resour., 24, 707–723, https://doi.org/10.1016/S0309-1708(01)00005-7, 2001.

Okoli, K., Mazzoleni, M., Breinl, K., and Di Baldassarre, G.: A systematic comparison of statistical and hydrological methods for design flood estimation, Hydrol. Res., 50, 1665–1678, https://doi.org/10.2166/nh.2019.188, 2019.

Pathiraja, S., Westra, S., and Sharma, A.: Why continuous simulation? The role of antecedent moisture in design flood esti-mation, Water Resour. Res., 48, W06534, https://doi.org/10.1029/2011WR010997, 2012.

Porporato, A., Daly, E., and Rodriguez-Iturbe, I.: Soil water balance and ecosystem response to climate change, The Ameri-can naturalist, 164, 625–632, 2004.

Rogger, M., Kohl, B., Pirkl, H., Viglione, A., Komma, J., Kirnbauer, R., Merz, R., and Blöschl, G.: Runoff models and flood frequency statistics for design flood estimation in Austria – Do they tell a consistent story?, J. Hydrol., 456-457, 30–43, https://doi.org/10.1016/j.jhydrol.2012.05.068, 2012.

World Meteorological Organization: Manual on Estimation of Probable Maximum Precipitation (PMP), WMO Publ., 1045, 291 pp., 2009.

Zorzetto, E., Botter, G., and Marani, M.: On the emergence of rainfall extremes from ordinary events, Geophysical research letters, 43, 8076–8082, https://doi.org/10.1002/2016GL069445, 2016.