Response to Review

We thank both referees for their reviews and comments. Below, we respond to each comment in detail. Referee comments are shown in black, our responses are shown in blue.

Referee #1

R1.1 I read and reviewed the manuscript "Leveraging reforecasts for flood estimation with long continuous simulation: a proof-of-concept study". While the idea of the research is familiar (I have guided students on the same topic), and i understand the difficulties that play when applying such research in the Alpine domain. I find that the manuscript has too many messages which makes the end results to scattered. It would be wise to focus on one topic (e.g. meteorology) and leave out maybe some things to make the manuscript better stick (e.g. dynamic downscaling, hydrological modelling, etc. I leave this to the authors) . In the manuscript, one is often referred to the Supplement which is not a part of the manuscript. The manuscript could benefit from an additional experimental setup section where the different experiments that have been conducted to answer the research question are described. The research question is not clearly defined in my view. The main issue with this type of work is if the data used is homogeneous and second if so, how the fusion of the trace is conducted and how this may affect results (effects of data assimilation in the first couple of days of the reforecast etc, leave out the first days of the RFs, etc). I didn't see that information in the manuscript.

- *Too many messages, research question*: We agree that the manuscript spans several components along the modelling chain. This is intentional, as the study is explicitly framed as a proof of concept, investigating whether, how, and under which limitations reforecasts (RFs) can be leveraged for flood estimation using long continuous simulation (CS). The focus is on very rare floods in an Alpine setting, where observations are scarce and information on flood frequencies is particularly limited. Establishing this proof of concept requires following the full chain from raw RFs through post-processing and temporal disaggregation to hydrological simulation, hydrological routing and flood frequency analysis. Owing to the Alpine setting, this is carried out at an hourly time step throughout the modelling chain, which, to our knowledge, has not been achieved previously using RF data over such long time series. Juxtaposing the RF-based approach with results from a stochastic weather generator – taken from the framework established in Viviroli et al. (2022) – further provides valuable context for estimating very rare floods, for which alternative approaches are scarce.

  That said, we acknowledge the referee's concern that the presentation could be streamlined and the core research questions – in terms of feasibility, consistency, and added value of the approach – could be articulated more clearly.

  We will therefore revise the manuscript to more clearly highlight the following questions:
  1) the feasibility of processing RFs to hourly resolution in a challenging Alpine setting;
  2) the feasibility of concatenating the RFs into very long time series that yield consistent hydrological simulations suitable for long CS;
  3) the extent to which flood frequency estimates derived from such RF-based CS differ from and complement estimates obtained using an established stochastic weather generator framework; and
  4) the identification of specific limitations of the RF-based approach within a long CS framework.

Moreover, we will move the section on dynamical downscaling to the Supplement to remove potentially excessive methodological detail from the main text. We will discuss this topic more concisely in a short paragraph within the Limitations Section (see our response to comment R1.9 for more detail).

- *Supplement*: Please see our response to comment R1.6.
- *Experimental setup*: We will clarify in the manuscript that the model chain setup is identical to that used in Viviroli et al. (2022), with the only change being the meteorological forcing, which is switched from a stochastic weather generator to RFs. Given this, we believe that a separate experimental setup section is not necessary.
- *Homogeneity*: Please see our response to comment R1.2.
- *Effects of data assimilation*: Please see our response to comment R1.3.

R1.2

Did you check the data is homogeneous (likely it is not if you used all these different IFS cycles)? I didn't see such an analysis

We agree that homogeneity of the reforecast data is an important issue when combining RFs from different IFS cycles. While we did not perform a formal test, a visual examination already revealed inconsistencies, and we accounted for them in the processing. This is documented at several places in the manuscript:

- *Sect. 2.2 (Reforecast data):* We visualized RF precipitation and temperature by forecast lead time and by year of initialization. This analysis revealed systematic inconsistencies affecting days 1–15 and year 2015, which led us to exclude these data from further processing. This assessment is illustrated in Fig. S1 in the Supplement, to which we refer from Section 2.2. Please see also our response to comment R1.3 below.
- *Sect. 3.1.1 (Bias adjustment):* We explicitly note that the bias structure of temperature and precipitation varies by year of initialization, reflecting changes between IFS cycles. To account for a major change in IFS behaviour, bias adjustment was performed separately for the periods 2009–2014 and 2016–2020.
- *Sect. 3.1.4 (Concatenation):* When constructing yearly time series, individual RFs were drawn only from the same initialization year. This restriction was introduced to minimize mixing of RFs originating from different IFS cycles within a single concatenated year.

Taken together, these steps aim to identify and mitigate the most relevant sources of inhomogeneity arising from IFS model evolution, and changing bias characteristics. We will clarify in the manuscript that minor residual inhomogeneities persist and cannot be fully eliminated.

R1.3

Did you discard the first days or the forecast as they are affected by data assimilation? And if not did you check if this influenced the outcomes

Yes. As stated in Sect. 2.2 (Reforecast data) (L104–106), we discarded the first 15 forecast days and used only data from day 16 onward. This choice was motivated primarily by the need to ensure independence between individual reforecast ensemble members by allowing sufficiently long development from initial conditions. Following Mahlstein et al. (2019), discarding days 1–15 should effectively eliminate any effects of initial conditions that affect the first days of the reforecasts. This choice is further mentioned already in Sect. 5.5 (Limitations) (L473–476), where we note that we only used the Extended Range Model Climate data (days 16–45) due to concerns about both the

independence of short forecasting times from observed weather and inconsistencies between the Model Climate (day 1—15) and the Extended Range Model Climate.

To improve clarity, we will emphasise this more explicitly in the manuscript as follows (additions in boldface):

- "To address the first issue, we **discarded the first 15 days and** used only data from day 16 onward [...]" (Sect. 2.2)
- "However, due to inconsistencies in the data and concerns about the independence of short forecasting times from observed weather, only the Extended Range Model Climate data were used (see Sect. 2.2), **effectively discarding days 1–15**." (Sect. 5.5; Sect. 5.4 in revised manuscript)

R1.4

Mas et al 2023 is grey literature can you supply doi / link as it doesn't seemto be available for others, supply a better reference which is available

The relevant material from Mas et al. (2023) has in the meantime become available in a publicly accessible project report, namely Viviroli et al. (2025). We will replace the two cited instances accordingly in the revised manuscript.

R1.5

Figure 7 colour blind friendly? Difficult to see the dots maybe for some

We thank the referee for raising this point. We checked Fig. 7 using two colour-blindness simulation tools (`https://bioapps.byu.edu/colorblind_image_tester`; `https://www.color-blindness.com/coblis-color-blindness-simulator`) and confirmed that the figure remains interpretable.

R1.6

Too often referred to results in Supplement materials in my opinion

We will consolidate all references to the Supplement to (a) avoid disrupting the flow of the main text and (b) implicitly clarify that the Supplement provides additional material that is not critical for the main text but remains relevant for readers interested in more details.

With this in mind, we have carefully reviewed all references to supplementary figures and will consolidate them accordingly for Figs. S1–S19 and S22–S23, typically moving the reference to the end of a paragraph and making it less prominent, while adjusting the text where necessary. However, since we consider the hydrological validation of the reforecasts an important part of our study within the refined research questions, we will move the corresponding two figures from the Supplement into the main text. (i.e., Figs. S20 and S21 becoming Figs. 8 and 9). This will result in twelve figures in the main manuscript, which we consider still reasonable. However, we remain open to keeping these figures in the Supplement while applying the same consolidation of references as described above.

R1.7

Table 2 Is not mentioned in text

Tab. 2 is actually referred to in the manuscript at line 90 and at line 96, where the reforecasts are introduced in detail. As with Tab. 3, the placement of Tab. 2 is determined by the journal's LaTeX template (see also our response to comment R1.8 below).

R1.8

Table 3 is not mentioned in text (and at strange position)???

> Tab. 3 is actually referred to in the manuscript text (line 361) and in the caption of Fig. 9. We agree that its placement in the current layout may not be optimal. The positioning of Tab. 3 is determined by the journal's LaTeX template, and we expect that its final placement will be adjusted by the journal during copy-editing in case of publication.

R1.9

Leave out the paragraphs on dynamic downscaling read like a research report

> We agree that the current subsection on dynamical downscaling provides extensive methodological detail that may not be suited for the main manuscript. To streamline the text, we will move a slightly revised version of this subsection to the Supplement. In the main manuscript, we will instead slightly expand the discussion of dynamical downscaling in the Limitations section (Sect. 5.3 in revised manuscript) and refer to the Supplement for further detail. We will also adjust the end of the Introduction section accordingly.

R1.10

Figure 10 why compare against the world? distracts

> We understand the referee's concern. The global envelope curve is included to provide an upper-bound context for extreme floods, as it represents observations from regions with substantially wetter climates and different hydro-climatic regimes. While the Swiss and European envelope curves are more relevant geographically and climatologically, the global curve helps to contextualise the magnitude of the simulated extremes relative to what has been observed worldwide.
>
> That said, we agree that the visual emphasis should remain on the regional comparisons. We will therefore reduce the visual prominence of the global envelope curve by using a lighter colour, so that it still provides context while being less visually distracting, and increase visual prominence of the Swiss envelope curve by plotting it in black.
>
> In addition, we will revise Sect. 5.2 (Envelope curves for floods) to explicitly mention the purpose of the global envelope curve and to highlight its relevance for the Ticino region, in particular for Maggia–Locarno, where RF-derived maximum floods approach the world record curve.

R1.11

I would also expect return plots related to temperature/snow melt

> We agree that temperature- and snowmelt-related processes can be relevant contributors to flood generation. However, in the set of catchments examined here, the annual maximum flood predominantly occurs during the summer months (see Sect. 4.2 on Hydrological validation). In these months, rainfall-dominated (convective) events prevail, and snowmelt contributions are typically small (see Viviroli et al., 2025). We therefore argue that a detailed investigation of temperature- and snowmelt-related return aspects is beyond the scope of this proof-of-concept study, and will mention this as a potential direction for future research. In this context, we also note that the role of snowmelt in very rare flood events has been investigated in more detail by Staudinger et al. (2025) on the basis of continuous simulations. We will add a corresponding text to the Conclusions Section.

R1.12

Why was TabsD not used for bias correction temperature? similar to RHiresD?

> The reason is that, in contrast to precipitation, it was not necessary to apply a two-step procedure for temperature.
>
> Precipitation is a spatially and temporally highly variable stochastic process. Deterministic quantile mapping cannot generate the resulting sub-grid stochastic variability, but only the systematic variations, such as caused by topography. It is well established that using quantile mapping for downscaling from model grid to station scale may therefore introduce severe artefacts, such as an overestimation of area-averaged precipitation extremes (Maraun, 2013; Maraun et al., 2017). This issue is particularly relevant when generating meteorological input for a distributed hydrological model, as is the case in our study. For this reason, we applied a two-step approach for precipitation (Volosciuk et al., 2017; Switanek et al., 2022): model output was first bias adjusted against observations at the grid scale (RHiresD) and then stochastically downscaled to the station scale.
>
> Temperature, in contrast, behaves much smoother both in space and time and shows much weaker stochastic sub-grid variability. Direct bias adjustment of modelled temperature against station values is therefore justified and does not suffer from the same problems as in the case of precipitation. Consequently, TabsD was not used.

R1.13

Figure 8 make the observations more visible (black?)

> We tested plotting the observations in black, but found that this did not improve the overall readability. In some cases, two observational series are shown, which requires the use of two distinct colour shades. It is more difficult to achieve this clearly using black and grey tones (given the grey grid lines) than it is with coloured symbols.
>
> To improve visibility while maintaining clarity, we will instead slightly increase the symbol size of the observations in Fig. 8.

R1.14

Conclusions are not very informative (what were the specific research questions?)

> We will revise the Conclusions to explicitly restate the key research questions (see comment R1.1) and summarise the corresponding findings. In addition, we will expand the Conclusions to explicitly discuss the role of snowmelt (see comment R1.11) and the uncertainty in the underlying weather model (see comment R2.1).

R1.15

Figure 9 colourblind friendly?

> As with Fig. 7, we have checked using two colour-blindness simulation tools and confirmed that the figure remains interpretable.

R1.16

Make data/code available in a repository

Due to the very large volume of data generated in this study (in the order of 250 GB), as well as the complexity of the data organisation – which is tailored to high-performance computing workflows in different research groups – we have refrained from making datasets available in a public repository.

Similarly, the code used for processing, post-processing, modelling and evaluation consists of multiple partly interdependent components that have been developed and adapted over an extended period. Preparing these workflows for public release would require substantial effort for curation, documentation, and generalisation that is beyond the resources available for the present study.

Efforts to curate selected datasets into a public database have been initiated together with the federal offices funding this study; however, due to the substantial effort involved, this resource will not be available in the near term. As stated in the manuscript, however, data and code used in this study will be made available upon reasonable request.

R1.17

Maybe focus on own results vs observations first before comparing against GWEX etc

We thank the referee for this suggestion. The integrated presentation of information from reforecasts, observations, control runs, and weather generator is intended to facilitate direct comparison across datasets. Nevertheless, we agree that the presentation can be clarified. We carefully went through the manuscript and have identified three sections where this comment is relevant:

- Sect. 4.3 (Flood estimation results): We will revise the section to more clearly introduce the sequence of comparisons and then consistently follow aforementioned sequence, namely RF-derived Annual Maximum Floods (AMFs) – observed AMFs – control run AMFs – GWEX-derived AMFs. This should clarify while preserving the integrated interpretation. Separating the presentations of datasets beyond this would, in our view, make the section unnecessarily fragmented and harder to follow.
- Sect. 5.1 (Comparison of reforecasts and weather generator precipitation): We will revise the paragraph that explains the context for the juxtaposition to first mention the reforecasts.
- Sect. 5.2 (Envelope curves): Here, the sequence of comparisons already follows the suggested order, no changes required.

Referee #2

R2.1

The paper is very well written and provides insight into novel methods to estimate extreme values for floods.

I believe the manuscript overstates uncertainty reduction by implicitly treating the 10 000-year RF dataset as a large set of independent samples. In reality, these are model-generated statistical draws from a numerical weather prediction system, sharing common physics, parameterizations, and structural errors. Increasing the number of RF realizations reduces Monte Carlo variability within that model, but does not reduce epistemic or structural uncertainty.

Uncertainty in the underlying weather model propagates directly into the flood estimates and this should be discussed. Systematic biases in precipitation extremes, storm persistence, and temperature–precipitation co-variability are inherited by the hydrological simulations and cannot be mitigated by ensemble size alone. Bias correction and stochastic downscaling adjust marginal statistics but do not correct errors in event dynamics, spatial coherence, or compound processes that control extreme floods.

As a result, the narrowing of exceedance curves reflects internal consistency of a fixed meteorological–hydrological modelling chain, not increased confidence in true flood return levels. The authors should explicitly state that RFs reduce sampling uncertainty conditional on one weather model and postprocessing framework, while leaving weather-model, structural, and climate-representation uncertainties largely unresolved.

> We thank the referee for their positive feedback and for the thoughtful remarks regarding uncertainty. We fully agree that these points are important to highlight. To address this, we will add a dedicated paragraph in the Limitations Section (Sect. 5.4 in revised manuscript). To maintain clarity in the expanded Limitations Section, we will also introduce a third level of numbering for better structure and readability. In addition, we will add a short paragraph at the end of the section on Flood return levels (Sect. 5.3 in revised manuscript) and re-state this point in the Conclusions.

Further Changes

It has come to our attention that at least three of the annual maximum flood measurements (1910, 1965, and 1977) conducted by the Federal Office for the Environment at Thur-Andelfingen are notably affected by bank overflow, effectively leading to an underestimation of peak flow, and that estimates of the unattenuated peak flows are available (see Scherrer et al., 2011; Hunziker, Zarn & Partner, 2017). This is relevant because the river channel was modified in 1998 to accommodate higher peak discharges, and our model chain represents this current state, in which bank overflow occurs less frequently. This also leads to better agreement between simulations and observations at this site. We will add adjusted peak flow statistics to the corresponding panel in Fig. 8 and update Sect. 4.3 (Flood estimation results) accordingly.

References

Hunziker, Zarn & Partner: Gefahrenkartierung Naturgefahren Thur: Technischer Bericht, 2017.

Mahlstein, I., Bhend, J., Spirig, C., and Martius, O.: Developing an Automated Medium-Range Flood Awareness System for Switzerland Based on Probabilistic Forecasts of Integrated Water Vapor Fluxes, Weather and Forecasting, 34, 1759–1776, https://doi.org/10.1175/WAF-D-18-0189.1, 2019.

Maraun, D.: Bias Correction, Quantile Mapping, and Downscaling: Revisiting the Inflation Issue, Journal of Climate, 26, 2137–2143, https://doi.org/10.1175/JCLI-D-12-00821.1, 2013.

Maraun, D., Shepherd, T. G., Widmann, M., Zappa, G., Walton, D., Gutiérrez, J. M., Hagemann, S., Richter, I., Soares, P. M. M., Hall, A., and Mearns, L. O.: Towards process-informed bias correction of climate change simulations, Nature Climate Change, 7, 764–773, https://doi.org/10.1038/nclimate3418, 2017.

Mas, A., Evin, G., and Hingray, B.: Simulation of MAT and MAP scenarios using the weather generators GWEX and SCAMP: EXCH Final working report – Large catchments, 2023.

Scherrer, S., Frauchiger, R., Näf, D., and Scheible, G.: Historische Hochwasser: Weshalb der Blick zurück ein Fortschritt bei Hochwasserabschätzungen ist, wasser, energie, luft, 103, 7–13, 2011.

Staudinger, M., Kauzlaric, M., Mas, A., Evin, G., Hingray, B., and Viviroli, D.: The role of antecedent conditions in translating precipitation events into extreme floods at the catchment scale and in a large-basin context, Natural Hazards and Earth System Sciences, 25, 247–265, https://doi.org/10.5194/nhess-25-247-2025, 2025.

Switanek, M., Maraun, D., and Bevacqua, E.: Stochastic downscaling of gridded precipitation to spatially coherent subgrid precipitation fields using a transformed Gaussian model, International Journal of Climatology, 42, 6126–6147, https://doi.org/10.1002/joc.7581, 2022.

Viviroli, D., Sikorska-Senoner, A. E., Evin, G., Staudinger, M., Kauzlaric, M., Chardon, J., Favre, A.-C., Hingray, B., Nicolet, G., Raynaud, D., Seibert, J., Weingartner, R., and Whealton, C.: Comprehensive space-time hydrometeorological simulations for estimating very rare floods at multiple sites in a large river basin, Natural Hazards and Earth System Sciences, 22, 2891–2920, https://doi.org/10.5194/nhess-22-2891-2022, 2022.

Viviroli, D., Staudinger, M., and Kauzlaric, M.: Extreme Floods in Switzerland. Hydrological scenarios for large catchments. Project report commissioned by the Federal Office for the Environment (FOEN) and the Swiss Federal Office of Energy (SFOE), https://www.bafu.admin.ch/de/projekt-extremhochwasser-schweiz-exch, 2025.

Volosciuk, C., Maraun, D., Vrac, M., and Widmann, M.: A combined statistical bias correction and stochastic downscaling method for precipitation, Hydrology and Earth System Sciences, 21, 1693–1719, https://doi.org/10.5194/hess-21-1693-2017, 2017.