Reviewer 2

This manuscript investigates the potential changes in flash flood frequency and magnitude to be expected in the Alpine regions following the ongoing climate change. The study setup is based on convection-permitting model simulations that are fed to a hydrological model.

The ability of this chain to reproduce flash flooding in the past is evaluated and the performance of this setup is compared with the one obtained by feeding global-scale renalyses to the hydrological model.

The study investigates one of the key questions concerning precipitation-driven hazards under climate change and adopts state-of-the-art modelling elements that I deem adequate to the task. The authors put a huge amount of work in the study, and the manuscript is well written and clear.

At the same time, the simulation setup comes short at addressing some aspects that I believe are important in shaping the results. My concerns mainly revolve around the two points below. Addressing them may require abundant work, including additional analyses. I am not sure this can be done within a major revision.

(1) The assumption of trading space for time in the case of flash floods needs to be better motivated. Trading space for time is generally used for processes that are not scale-dependent. Flash floods are, as they can be generated by diverse meteorological forcing, in relation to the basin area and characteristics. This means that we can have nearby basins that respond to different meteorological forcing with different spatial and temporal scales. These different forcing can be represented with different accuracy by climate models.

Indeed from a hydrological perspective the processes are very scale and location dependent. From a meteorological point of view there is far more variation, except for the strong south-north influence of the Alps. The exact location of heavy rainfall events in the CPM runs is affected by the internal variability of the CPM and in the current and future climate runs the location of heavy rain events can shift between basins. Note, however we focus on flash floods in the summer-fall period which are less affected by frontal systems.

This aspect is neglected in the flash flood validation. For example, it is possible to have a peak in a larger nearby catchment 2 days apart from the original event due

to completely different meteorological forcing (an incoming large-scale front as opposed to isolated convection, for instance).

On this aspect, how is 3 days chosen (see line 198)? It appears a long time to me, unless there are known uncertainties in the timing of the used databases.

See also reply below, because CPM simulations driven by ERAInterim are affected by internal variability storms may not appear at the right place/time. The three days is based on early investigations into these modelled flash flood event.

(2) Several processes in flash flood generation are non-linear. Therefore, quantitative values of many variables are critical for flash flood response. I feel that the model setup neglects the biases in the CPM simulations. Quantitative biases are expected to be enhanced by the non-linear hydrological processes typical of flash flood generation. While it is true that there a lot of attention in the validation of the modelling chain (section 2.6), the performance of said chain are not very high (e.g. see figure 3).

This a reply also provided to reviewer . The reviewer(s) attributes all performance issues of performance to the hydrological model used and does not take into account the quality of the rainfall forcing dataset used. We use dynamically downscaled ERAinterim reanalysis data to drive the hydrological model. This means that the climate model was forced with boundary conditions from ERAinterim and is not corrected by data assimilation to correct locations of pressure systems etc on the right spot. Therefore, the forcing for the hydrological model is affected by the internal variability of the climate model and rainfall systems may end up at the wrong spot or happen at the wrong time. This may not be clear from the manuscript and will be clarified more in a next version of the manuscript. We use this dataset not because this is the best forcing, but to be in line with the future climate model output that is also used forcing. This enables us to have a fair comparison between changes when we compare present to future climate runs. That is also the reason why, in addition to the ERAInterim forcing, we used ERA5 reanalysis data (still very coarse and not ideal in the Alps) which is completed controlled by the data assimilation to demonstrate that correspondence with observations will improve when better and with more data assimilation and higher resolution reanalysis data will be used. Given that our forcing is coming from a CPM driven by ERAinterim boundary conditions and is affected by internal variability we think Figure 2-4 shows that the hydrological model has credible performance in the Alps (see also Imhoff et al., 2020, Imhoff et al., 2024 and van Verseveld et al., 2024). We agree that the results in Figure 5 are not good but note that with this harsher criterium, as mentioned, dynamically downscaled ERAInterim data as forcing might play a major role here (poor boundary conditions CPM, wrong placement of storms, amounts, etc). We think the hydrological model set up as presented is credible (and one of its first to do this at this scale) for the conducted analysis. We were and are not aware of an alternative (open-source) model setup. And as

such we think the manuscript is a valuable contribution to the ongoing scientific debate/discourse on this topic.

- How much can we trust the modelling chain quantitatively? Is the relatively low performance related to deficiencies in the CPM simulations or in the hydrological model? Perhaps some of these answers can be replied to by exploring the simulated precipitation fields and comparing them to the triggering ones. For example, in the case of the database from Amponsah et al, radar estimates for all the events are provided.

We used ERA5 reanalysis data (relatively coarse and not ideal in the Alps), the successor of the ERAinterim dataset to demonstrate that correspondence with observed discharges is reasonable. Figure 2-4 show that the model has credible performance in the Alps (see also Imhoff et al., 2020, Imhoff et al., 2024 and van Verseveld et al., 2024). We agree that the results in Figure 5 are not good but note that with this harsher criterium, as mentioned, dynamically downscaled ERAInterim data as forcing might play a major role here (poor boundary conditions CPM, wrong placement of storms, amounts, etc). We think the hydrological model set up as presented is credible (and one of its first to do this at this scale) for the conducted analysis. We were and are not aware of an alternative (open-source) model setup. And as such we think the manuscript is a valuable contribution to the ongoing scientific debate/discourse on this topic.

- The text in lines 301-303 attempts an explanation on why there is no adjustment but, in light of the non-linear responses mentioned above, I believe prudence should be put in saying that comparing model with model decreases the importance of the biases.

Indeed no bias-correction was applied as there is no homogeneous observational datasets that covers the whole area. Besides that, the climate model simulation windows (~10 years) are too short for a reliable distribution match as for example required in quantile mapping. The periods are so short as the CPM compute is intensive and data amounts are massive. We will adapt the text stating less firmly that the influence of the bias is less relevant when comparing model with model outcomes.

- It is not clear to me how is the hydrological model calibrated. Is it calibrated based on the CPM simulations? ERA5? ERA-Interim? I believe the calibration strategy should be better described as it also plays an important role in shaping the results.

The hydrological model was not calibrated. We used the setup as described in Imhoff et al. 2020. We explored the sensitivity of the lateral hydraulic conductivity on simulated discharges as explained. Please see Imhoff et al 2020 and van Verseveld et al 2024 for

more information. Imhoff et al 2024 (in review) also provides insight on the performance. We will provide some more explanation in the adjusted manuscript.

Minor comments:

- Fig. 6,7: the reported variance is very large. It could be related to regional scale variability and scale-dependence (area) in the climate change response. The organisation into large-scale fluvial catchments may mask these aspects. Do you see any spatial pattern? Is it possible to organise it somehow into regions with homogeneous response? And/or by basin area?

Given that the datasets are short and therefore we adopted a space for time trade we don't think this is appropriate to do beyond what we already did in the manuscript.

- In Fig 6, I'd suggest using a log transformation on the y axis

Good suggestion. Will be adjusted in revised manuscript

W J Van Verseveld , A H Weerts , M Visser , J Buitink , R O Imhoff , H Boisgontier , L Bouaziz , D Eilander , M Hegnauer , C Ten Velden , B Russell Wflow_sbm v0.7.3, a spatially distributed hydrologic model: from global data to local applications Geoscientific Model Development (in press), 2024.

Imhoff, Ruben and Buitink, Joost and van Verseveld, Willem and Weerts, Albrecht, A Fast High Resolution Distributed Hydrological Model for Forecasting, Climate Scenarios and Digital Twin Applications Using Wflow_sbm. Submitted to EMS, available at http://dx.doi.org/10.2139/ssrn.4757726

Imhoff, R.O., van Verseveld, W.J., van Osnabrugge, B., Weerts, A.H., 2020. Scaling Point-Scale (Pedo)transfer Functions to Seamless Large-Domain Parameter Estimates for High-Resolution

Distributed Hydrologic Modeling: An Example for the Rhine River. Water Resources Research 56, e2019WR026807. doi:10.1029/2019WR026807.