Reviewer 1

This study investigates the potential changes in flash floods over the Alps due to climate change. The aim of this study was to evaluate the use of a convection-permitting regional climate model in combination with a distributed hydrological model to assess future changes in the frequency and magnitude of flash floods over the Alps. The UM convection-permitting model was used to project future climate at high spatial and temporal resolution. These projections were used as inputs to the distributed hydrological model wflow_sbm. The ability of the hydrological model to simulate historical flash floods was first assessed using reanalysis data. The modelling framework was then used to investigate future changes in flash floods.

Although the topic of the paper is very interesting and has not yet been addressed in the literature, I do not believe that the methodology is adequate to address the research question formulated in the introduction. The modelling framework and data set used in this study are not robust enough to support the conclusions.

Major comments:

1. The ability of the hydrological model to reproduce historical flash floods in the study area is very low. Apart from the fact that the methodology for assessing this is questionable (see comments below), the results show that the model is not able to detect floods based on the threshold approach for about half of the stations studied (Fig. 5). The hydrographs in Figure 3 also show the limited performance in simulating most floods (and not just flash floods). If the modelling framework does not capture the dominant processes triggering floods in this region in the historical period, it is very unlikely that it will do so under changing conditions. There could be several reasons for this (e.g. uncertainty in the input data). It could also be argued that a better modelling framework has not yet been developed. However, since the current modelling capabilities (at least for the modelling framework of this study) are not robust and reliable enough in this context, I do not think it should be used to study future changes in flash floods.

The reviewer attributes theall performance issues of performancemainly to the hydrological model used and does not take into accountonly briefly mentions 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 copleted controlled by the data assimilation, ERA5 is used 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 strorms, 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.

2. The definition of a flash flood lacks a very important feature: the temporal dynamics of the storm and/or flood event. To determine what is a flash flood, the authors use the specific peak discharge and the size of the upstream catchment. This definition was adapted from Amponsah et al. (2018) by removing the storm duration as a selection criterion (not mentioned in the manuscript). This means that any discharge exceeding the specified threshold and occurring in a catchment of less than 3000 km^2 is considered a flash flood, regardless of whether it was triggered by a uniform precipitation event lasting a few days or by a very intense and highly spatially variable convective storm event. Floods associated with slow catchment dynamics can theoretically be considered as flash floods in this modelling framework.

As mentioned in the manuscript, we use the definition of Amponsah et al (2018) to be able to compare our results- to their results of flood impacts (no other observational datasets are available). Any definition of flash floods will have problems or issues. Indeed, floods associated with slow catchment dynamics could theoretically be consisidisdered but as we focus in this work on the Alps in summer/fall in steep terrain, this is not very likely. Alternatively, we could alter the title of the manuscript to future changes in summer and fall flood frequency and magnitude over the European Alps.

3. The validation/evaluation methodology is not suitable for assessing the ability of the modelling framework to project future changes in flash floods.

    a. The aggregation to the daily time step for model validation does not allow to evaluate the ability of the model to simulate flash floods, because the hydrological processes involved affect the flow at sub-daily time steps.

       Unfortunately, there is no other way to do the validation as we do not have access to hourly discharge observations for the validation. The hydrological model accounts for sub-daily processes, we run at an hourly time-step.

       Formatted: Highlight

    b. Basing the performance assessment on peak flows and the KGE calculated over the entire time series is not sufficient to assess whether the modelling framework was able to capture the dominant flood generation processes.

        i. As the catchments do not necessarily only experience flash floods, the number of points related to these floods used to calculate an error criterion is small compared to the rest of the time series. Therefore, a high KGE value in this case does not mean that the model performs well in simulating flash floods. In addition, the KGE values are quite low for many stations (KGE< 0.6 for about ½ of the stations, see Table 3, Crochemore et al. 2015).

           In principle this is true, yet in these relatively small hilly / mountainous catchments we expect the majority of events to be flash floods. Indeed, there are quite some low KGE values. On the other hand, for about ½ of the stations we find KGE values higher than 0.6. Improving model performance in these type of small and fast responding basins is quite challenging.

        ii. Peak flow analyses rely on single points, which can be highly uncertain.

           Indeed, and as said above that also hampers the validation, but this would hamper the validation of any type of modelling framework.

    c. As part of the validation/evaluation analysis is based on nine historical floods, it would have been possible to plot the flood hydrographs and compare the simulations with the observations (if available).

       Yes this can be added in the Supplement

4. The paper lacks a comprehensive description of the streamflow dataset, in particular why this dataset is suitable for answering the research question. It is not clear why only 130 gauged stations were used in this study and how they were selected.
    a. Were the stations selected because their catchments experience flash floods? Were they chosen because of data availability, low human influence...?

       The stations were indeed chosen because they were merely located upstream in the basins at locations, upstream of lakes / reservoirs where the human influence was still limited. For some basins, especially the Rhine, we had access to a large number of stations. Whereas for the Adige and Po it was hard to get data. We will extend the manuscript with an explanation in the revised version.

    b. What are the hydrological processes affecting river flows in these catchments?

       This will be added to the revised manuscript. Most relevant processes are (saturation excess) surface runoff, sub-surface flow, snow melt and baseflow from the groundwater. The slopes in the upper parts of the sub-catchment have a steep gradient, resulting in fast subsurface and surface runoff. The permeable soil-layers are of limited depth often followed by underlaying rocky soil types allowing for limited infiltration.

    c. Are all the catchments and sub-catchments prone to flash floods and how often compared to the other floods?

       We didn't conduct such an analysis as the area is quite big

    d. Why were stations with different temporal resolutions chosen?

       This dependents on data availability. For those stations where we had access to hourly data we used hourly data. For other stations we had to use daily data. This comment will be added to the manuscript.

    e. A large part of the study area is not covered by gauged stations.

       That is correct, we did not have access to other station data. Yet, we assume that model performance in neighboring sub-catchments will be of comparable quality.

5. No bias correction was applied to the climate projections. This point is discussed in Section 4, the main arguments being that bias correction can distort the change signal and that the observational data available over the

study region do not have the resolution of the CPM. However, most hydrological impact studies apply some form of bias correction, as biases are usually quite large at the hydrological scale (e.g. Teutschbein and Seibert, 2012). It is not clear how the choice not to apply a bias correction affects the results of this study, as there is no assessment of the climate variables for the historical period compared to the observations. Are the biases large enough to significantly change peak discharge simulations?

The main reason not to apply a bias correction is that there is no homogeneous observational datasets that covers the whole area. Unfortunately, there is not sufficient sub-daily / hourly data available for bias-correction over the full area. In the paper of Ban et al. (2021) the data was validated. Their analysis showed that indeed biases were present in the extreme rainfall amounts. We will add an additional comment to the discussion that indeed this may affect the simulation of flash floods. Besides that, the climate model simulation windows (~10 years) are short as the compute is intensive and data amounts are massive. Therefore, we decided not to perform any bias correction as explained in the manuscript and look at the differences between the current (downscaled ERAInterim) and future climate runs.

6. An analysis of changes in flood drivers might have been expected. Other studies reporting on the potential drying of the region are mentioned (l276 to 280), but the states and fluxes of the hydrological model could have been used for a more in-depth analysis.

We agree that this could have been done. However, we didn't store all the output variables of the hydrological models

Formatted: Indent: First line: 0.49"

7. As I understand it, one of the objectives of this study was to assess the potential added value of using a CPM in combination with a distributed hydrological model to assess future changes in flash floods over the Alps. In order to assess the added value of such a framework, this study should have included a benchmark, such as the projections of a regional climate model at 0.11° resolution.

We disagree as we don't see the additional benefit of such an analysis in this manuscript ~~(this was already done in Ban et al 2018) here. For this manuscript, it would distract from the main message.~~

This comparison between CP-RCM and RCM is already conducted for the rainfall extremes (see Ban et al 2018) *driving the flash floods and was not specifically reconsidered in this study.*

Formatted: Space Before: Auto

Other comments

8. Some of the figures and tables do not highlight the results well enough (suggestions for improving the figures below).

    a. 2: add the delineation of the study region

Yes we will do this

    b. 3: one hydrograph per line; extend in width; eventually show a smaller period.

Yes we will adjust the plot

    c. 5: increase resolution; change “f1” and “peirce_ss” to “F1” and “Peirce”. Add number of stations in the legend. Add number of stations with skill score = 0.

Yes we will adapt the plot

    d. 2: add a column with observed peak specific discharge.

    Not sure if this is available in the datasets mentioned but we will have a look

    e. 6: apply a transformation on the values of the x-axis to improve visualisation.

We tried but this was in the end the best choice in our view

9. L11-12: “and to determine a suitable threshold definition for flash flooding.” I did not understand why the definition of a threshold for flash floods is one of the objectives and why it is based on modelling results and not on observations.

We will rephrase the abstract as this was not the objective and we used a definition based on Amponsah et al (2018).

This *will be further explained in the revised manuscript. There is a bias in flows between observations and simulations. To make a reliable comparison between flash flood occurrence modelled for the current and future climate we need a threshold that is based on the historic modelling and apply this to the future simulation results as well to detect future flash floods.*

10. L148-156: why is glacier modelling mentioned in these lines? The authors show that glaciers do not have a significant influence on the occurrence of flash floods, but do not conclude on the implications for their modelling framework.

The wflow_sbm models were setup with glaciers as mentioned. We i investigated possible effects (none found) on simulations results as the m odels are run for current and future climate

They are part of the wflow_sbm model and are therefore included in the description. Glacier melt is a slow process mainly occurring in the middle and end of summer, its contribution to flash floods is assumed to be neglectable. The last remark will be added to the manuscript.

11. L 198: the regional approach to flash flood validation should be better explained, e.g. with a simple figure showing how the threshold approach is used to determine whether the model detects a flash flood or not.

This will be added

12. L205: "rare extreme events". If the flash floods considered in this study are rare events, the limited duration of the CPM projections (10 years) may not allow such events to be studied. I would have expected a small analysis of the flood events that occurred in the historical period for the 130 stations.

This analysis was conducted for historical period only and not the CPM p projections

Indeed, this is a severe limitation, that is why we introduced the concept of trading space for time. We do already discuss this in the discussion, but will extend this part.

13. L 268: "5.6m3s−1km−2 compared to 4.49m3s−1km−2". The difference between the two values is most likely within the uncertainty range of the hydrological model (to be confirmed with the observed peak values to be added to Table 2).

This is an interesting comment however we are not sure at this point, and this will investigate this when making the plots/adjust the table.

14. L276 to 280 could be moved to the discussion section.

We will do this

References:

Louise Crochemore, Charles Perrin, Vazken Andréassian, Uwe Ehret, Simon P. Seibert, Salvatore Grimaldi, Hoshin Gupta & Jean-Emmanuel Paturel (2015) Comparing expert judgement and numerical criteria for hydrograph evaluation, Hydrological Sciences Journal, 60:3, 402-423, DOI: 10.1080/02626667.2014.903331

Teutschbein, C., & Seibert, J. (2012) Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. Journal of hydrology, 456, 12-29, DOI: 10.1016/j.jhydrol.2012.05.052

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.

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).

~~*We need to extend this with a comment on the falsh flood trading space for time especially with ERA5 forcing.*~~

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.

(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).

- 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

Reviewer 3

The manuscript submitted by Zander et al. presents the results of a study aiming at quantifying changes in flash flood frequency in the Alpine domain related to "worst case" climate change (RCP 8.5). The authors make use of the results of recent convention-permitting regional climate models' (CP-RCMs) hourly simulations with a 2.2 km spatial resolution and feed these to a distributed hydrological model set up for the five largest basins outspringing in the central European Alps, with a cell size of roughly 1 km^2.

The hydrological model is not calibrated, if I understand it correctly, but parameters are rather inferred from (pedo)transfer functions. The modeling chain is first validated and evaluated using common/already existing global reanalysis products (daily ERA5 and hourly *CP-RCM-downscaled* ERA-Interim) as inputs, and continuous discharge measurements together with peak specific discharge events extracted from two flash flood data bases (HANZE and EuroMedeFF) as observations. In a second step changes in flash flood frequency are investigated, inferred from the simulations performed using the raw CP-RCMs data (i.e. no downscaling resp. correction is undertaken for CP-RCMs data, but a bare remapping to the resolution of the hydrological model is performed).

The manuscript as a whole is well structured, the methods are generally understandable or mostly supported by relevant sources, the discussion is good, but quite (too?) short. I think some parts of the latter rather belong to the methods part (e.g. the choice of not bias correcting the data). In general, some important assumptions/concepts/choices/etc.., as well as the novelty of the study could (should?) be better elaborated and highlighted.

I am not a native speaker, and as such not the best judge, but I dare to say the paper is mostly written correctly, and it mostly reads fine, just sometimes is not consistent in the terminology (e.g. Current climate and Historical climate referring to the same dataset).

The quality of the figures featured in the manuscript could be definitely improved, while the relative captions accompanying these are clear and provide good descriptions.

Despite the main idea of the study is in fact really interesting and to my knowledge has not yet been addressed with such a framework, I have several concerns about some assumptions made and the implementation of this analysis:

1) The hypothesis of trading space for time: flash floods are a specific subcategory of floods. A flash flood is usually a quite localized (10-100 km^2), short-lasting event (min to hr), in order to have a more widespread effect and

have a regional character it needs some "ingredients", such as special weather patterns (for the European Alps probably an atmospheric river or slow-moving storms) and/or special preconditions, such as saturated soils. The time window of potential occurrence and the time steps used in this study for validation/evaluation are way too large for this kind of floods (3 days and hourly data aggregated to mean daily discharge, respectively), and don't really allow to make any conclusion on the goodness and the suitability of the modelling framework applied here.

We understand that the definition of what is a flash flood can be quite sensitive. However, we followed the definitions in the impact papers e.g Amponsah et al 2018

If we need to adjust the title of the manuscript, we are happy to do so as stated also in replies to reviewer 1 and 2. We did follow this definition to be able to make a connection to observed impacts

2) The definition of flash flood used in fact completely lacks the inclusion of the time component, it is highly questionable if this assumption can be justified.

See reply above

3) The lack of the observed specific discharges and of the location of the flash floods for the nine recorded flash floods used in the validation: In Table 2 only modelled peak discharges are reported, and often the whole subbasin is provided as the region being affected, even though more precise indications are available in the data bases used. E.g. the events on the 6th of June and on the 8th of September in the Rhone subbasin: both events took place in France (6th of June in Isère and 8th of September in Ardèche and further downstream) considerably distant from the outlet of the Rhone subbasin actually modelled (Rhone down to Geneva), for which only two gauging stations are used. The swiss part of the Rhone in 2002 experienced only locally some relevant floods in November, as well as the Ticino, to my knowledge. Or can the authors prove better?

We used the database as it exists. In that database of impacts not much more information is provided.

4) Discharge data: the discharge data are spatially extremely biased. Why is that? There are considerably more stations available also for the Rhine, the Rhone and the Po. In the manuscript there is no description of how the choice of the stations was made, and also a more informative overview of the stations used is missing (e.g. with the size of the subcatchment, the length of

observations, the river and the corresponding basin). This could be easily provided also in Appendix or as Support Material.

We used readily/publicly available datasets. That requested information is provided in the datasets mentioned.

5) The F1 and Peirce skill scores are actually quite poor. The reason and the dynamics behind these results are not really further investigated, so that is difficult to understand and trust, what the model is doing.

This ~~a~~ reply is also provided to reviewer 1. 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 bou~~n~~ndary 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 ~~copleted~~ 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 strorms, 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.

6) 10 years reference time frame for inferring changes in frequency: is really a decade a period long enough to make any inference on changes in frequency of flashfloods as compared to the current period? (this is part of the original hypothesis of trading space for time) The authors cite studies on extreme

value statics from short time series of rainfall, but floods and in particular flash floods occur significantly less frequently than heavy precipitation (see e.g. Modrick and Georgakakos, 2015). One might also argue, that the results might be strongly influenced by the chosen decade, depending if it falls within a flood-rich or a flood-poor period (see also Lun at al.2020, Fischer et al. 2023). A discussion and a climatological embedding are missing.

We will extend the discussion addressing the comments above. We agree with the fact that with the decade of data it is nearly impossible to conduct flood analysis. This is the reason why we introduced the concept of trading space for time. CP-RCM simulations are computationally very demanding due to the – for climate models – relatively high spatial and temporal resolution. We could therefore only get access to 10-years of data for the current and future climate – and this indeed set serious limitations to the study.

7) No bias correction/ downscaling of climate data: while it might be true that applying a correction is not always necessary when switching meteorological product (see Reed et al. 2007 and Alfieri&Thielen, 2015), it is however difficult to judge in this study if this is a valid choice, because there is no evaluation from the climatological-hydrological point of view of the "historical" period (current period). Objectively, we only know from Ban et al.2021, Fig.7 that in France and Italy summer bias in frequency of hourly precipitation is larger than -25%, while in Switzerland this bias is between -5 and -25%, and in fall is between +25-40% when downscaling ERA-Interim with UM.

We mention this limitation quite clearly in the manuscript. There is/was no suitable homogeneous hourly rainfall dataset at the time the study was conducted. More over the periods are rather short. We believe the work and setup is credible and contributed to the ongoing discourse/debate regarding impacts of climate change especially over the Alps. See also our replies to reviewer 1 and 2 regarding the same point. The data from the experiments from the EUCP project have limitations for climate impact modelling, but with these limitations as stated in the manuscript we believe we make a valid contribution to the scientific debate.

Because of these considerations, I think the manuscript requires substantial revisions and additional work, before it can be considered for publication.

Please find my specific and technical comments here following.

Specific comments

Abstract:

- Please be consistent in the terminology, either you always write Flash Floods with two capital letters (as the first two words in the abstract), or you stick to flash floods (then make the F in floods small)

Thanks we will follow the suggestion and use flash floods throughout the manuscript

Introduction:

- P2-L27: Kotlarski et al.(2023) is not listed in the References..

Thanks we will add this reference

- I miss an embedding in what is currently into place on a European level for simulating and detecting flash floods. Being one of the deadliest flood (hazard) type, quite some things and studies were performed on the continental scale in the last decade. E.g. https://www.efas.eu/en/flash-flood-indicators or Ritter al al. 2021?

We will extend the embedding in the introduction and discussion. Yet, most of these studies focused on historical conditions

Data and Method:

- P3-L87: what do you mean with ridge height?

~~I think t~~The 'divide' between the Northen part of the Alps and the Southern part of the Alps

- Being a swiss hydrologist and seeing your study area is basically covering the whole of hydrological Switzerland, and this makes probably also most of your study area (?), another useful database with peak specific discharge to be used when analyzing your flash floods on a regional scale could be found here https://opendata.swiss/de/dataset/grosse-hochwasserabfluesse-in-der-schweiz (however most of the explanations are available only in German, I am afraid). As I said already before, we miss an overview of the (numeric) areas of the basins and also of their sub-basins, it is important to know the spatial scale, in particular because you make use of area as a threshold and also of peak specific discharge as the main output variable.

The hydrological model wflow_sbm resolution is 30 arcsec (0.008333). Thanks for the link, that dataset was not known to us at the time of the study.

- 1: why do you use OpenStreetMap to produce this figure?(and also Fig. 2 later on) It would be easier (and nicer?) to produce these figures yourself.

Thanks we consider your suggestion

- P5-L119: Chan et al. (2020) is also missing in the References.

We will add Chan et al (2020)

- P5-L121: is a one in a year event an extreme event? I would reformulate this.

We will reformulate this

- P5-L127-129: I am not sure I understand why this is so not “unambiguous” (it has been done many times, simply assuming clear-sky conditions and solving a nice equation using latitude, day of the year –for estimating solar declination- and hour angle?). If I understand it correctly you assume there is radiation also in the night? I believe Osnabrugge et al. 2019 used hourly downward radiation.

We think the followed approach has very minor effect/impact on the results given the fact that we focus on flash floods. Correct Osnabrugge used observations of SSRD from LSA SAF combined with ERA5 values os SSRD.

- P6-L143: wouldn’t it be easier if you would simply say you used a priori PTFs and upscaling rules as Imhoff et al.(2020)? Most hydrologists would then know you used something similar to an MPR.

Thanks we will consider this remark

- P6-L147: Myeni et al.2015 is only pertinent for MODIS data, I think, for CORINE data please cite properly Copernicus.

Thanks, we will cite CORINE correctly

- P6-L160: why do you use ERA5 as forcing to perform the sensitivity analysis, and in general for model validation? I don’t understand why you take a daily dataset - kind of setting it as a benchmark- while all the rest is performed on an hourly basis. What is the rationale? Why didn’t you use ERA5 Land for example, which is available at fine spatial and temporal scales (hourly, about 9km), more similar to the meteorological products you used in the rest of the study?

See remarks reviewer 1 and 2 regarding the

- P6-161-162: why do you choose specifically these two stations specifically to conduct a sensitivity analysis on a factor that is applied to the whole modelling domain? The Thur river is nested in the catchment of the Rhine at Basel, which is per se not bad, but 1) why don't you perform this same analysis at other stations in different climatic regions? 2) The Rhine at Basel might be quite strongly influenced (and the sensitivity reduced?) by the several large regulated lakes upstream, did you consider it when making your choice?

- P6-L162: based on which criteria was a factor 100 defined as satisfactory?

Model performance/KGE see Appendix A1

- Please avoid saying downscaled while you actually simply remapped data, it is misleading (e.g. P7 L167)

It is downscaled uniformly – considering a lapse rate correction, we don't think this is misleading and we will clarify this more clearly

- P7-L176: why do you only use KGE for validation? I don't think it is appropriate to use only KGE as a goodness-of-fit measure for validating a modelling framework devoted to flash floods (in particular if you only use mean **daily** discharges!)

As mentioned above we could only validate against observed daily discharges as we had not access to sub-daily observations. The validation of the modelling framework is not the core of the study – we focus on the application for climate change and therefore considered one performance measure sufficient. In Imhoff et al we already paid a lot more attention to the validation of wflow_sbm.

Formatted: No bullets or numbering

- What is the spin-up period of the climatological models? And of the hydrological model? The spin-up period is not mentioned anywhere, I believe?

The spinup is two years for the evaluation and one year in for the current & future climate runs for the hydrological models. For the CPMs I refer to related papers of Ban et al (2018) or reports from the EUCP project.

- P7-L210: you sometimes switch “Historical” and “Current”, please be consistent throughout the manuscript (I would suggest to always use “Current” and remove the reference “Historical”).

Thanks we will make this consistent throughout the manuscript

Results:

- It would be very beneficial if you would provide more detailed information on the selected floods (i.e. extend Table 2), and show some hydrographs of how this flash floods were picked (or not) by the model.

We will this was also requested by the reviewers 1+2

- I would remove the grey background in all figures.

We will consider this when revising

- P9-L221-223: In Figure 3 you show one random year of simulation for two catchments, this is not really telling us too much about the model performance and ability to pick up the annual cycle or snowmelt overall. The Thur is a subalpine (transition) catchment (so snowmelt is relatively limited), and the Ticino in Bellinzona is quite strongly influenced by hydropower upstream (and the largest summer events are all missed?).

  We still think that showing the Thur is valuable for simulating floods. We will consider replacing the Ticino basin. These basins were only included for illustrative purposes.

- 7: maybe making additionally a plot per basin with subcatchment size on the x-axes would help to see if the very large values come e.g. from very small catchments? It is difficult to disentangle the actual meaning and consequences of the results presented in Fig.7. And also, if the results would look the same if e.g. precipitation intensity would be aggregated and plotted for the same basin and subcatchment subdivision.

We will consider this in the resvision.

Discussion:

- Did you consider that flash floods might be also shifting in time/season in the future?(see e.g. Blöschl et al. 2017 and Tarasova et al. 2023)

Ideally this would follow from the analysis conducted CPM-distributed hydrological model approach. We will make a remark about this in the discussion

- In general the discussion could be expanded..

Formatted: Indent: Left: 0.67"

Formatted: No bullets or numbering

Formatted: English (United States)

We will do this thanks for the suggestion

Technical corrections

- Throughout the text and in some caption you use "&", please avoid doing this and use the word "and" instead, which is more appropriate for a paper.

We will

- Please see the attached document for more technical corrections.

Thanks we will take them into account

References

Alfieri, L., Thielen, J. (2015): A European precipitation index for extreme rain-storm and flash flood early warning; *Meteorol. Appl*., 22(1), 3–13, https://doi.org/10.1002/met.1328

Blöschl, G. et al.,(2017): Changing climate shifts timing of European floods; *Science* **357**,588-590(2017);DOI:10.1126/science.aan2506

Fischer, S., Lun, D., Schumann A.H., Blöschl, G. (2023): Detecting flood-type-specific flood-rich and flood-poor periods in peaks-over-threshold series with application to Bavaria (Germany); Stochastic Environmental Research and Risk Assessment **37**, 1395 – 1413, https://doi.org/10.1007/s00477-022-02350-8

Lun, D., Fischer, S., Viglione, A., & Blöschl, G. (2020): Detecting Flood-Rich and Flood-Poor Periods in Annual Peak Discharges Across Europe; *Water Resources Research*, *56*(7). https://doi.org/10.1029/2019wr026575

Modrick, T.M., Georgakakos, K. P. (2015): The character and causes of flash flood occurrence changes in mountainous small basins of Southern California under projected climatic change; *Journal of Hydrology: Regional Studies*, Vol. 3, 312-336, ISSN 2214-5818, https://doi.org/10.1016/j.ejrh.2015.02.003

Reed, S., Schaake, J., Zhang, Z. (2007): A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. *J. Hydrol.* **337**: 402–420. https://doi.org/10.1016/j.jhydrol.2007.02.015

Ritter, J., Berenguer, M., Park, S., Sempere-Torres, D. (2021): Real-time assessment of flash flood impacts at pan-European scale: The ReAFFINE method; *Journal of Hydrology*, Vol. 603, Part C, 127022, ISSN 0022-1694, https://doi.org/10.1016/j.jhydrol.2021.127022

Tarasova, L., Lun, D., Merz, R., Blöschl, G., Basso, S., BErtola, M, Miniusii, A., Rakovec, O., Samaniego, L., Thober, S., Kumar, R. (2023): Shifts in flood generation processes exacerbate regional flood anomalies in Europe. *Commun Earth Environ* **4**, 49; https://doi.org/10.1038/s43247-023-00714-8