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 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 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.

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 condisdered 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.

    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

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 (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 driving the flash floods and was not specifically reconsidered in this study.

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

<u>References:</u>

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.