In the following, we provide detailed point-by-point answers to the reviewers' comments. Original comments are provided in blue, and our responses are shown in black.

The study by Czakay et al. applies an LSTM model, trained by ERA5 data, in combination with an ensemble of 31 downscaled CORDEX CMIP5 to investigate how climate change may impact rain-on-snow floods in Germany. The study is well-designed and comprehensive, and it provides some interesting new findings. Still, I have some questions and comments that should be considered in a revised manuscript.

We thank the reviewer for a positive evaluation of our work and for the insightful suggestions.

General comments

1. My major concern is about the detection of peak-over-threshold events and the definition of floods. The authors use a threshold value to derive on average five peak flow events per year (POT5). A flood is then defined as POT5 events exceeding a return period of two years. How is the independency of flood events guaranteed? Two subsequent peaks may actually belong to the same flood event. This could have major impact on the results and their interpretation

   We thank reviewer for an insightful comment and apologize for not providing details on the selection criteria.
   To ensure the independence of the peaks detected by POT5, we applied the criteria suggested by USWRC (1975), which states that two flood peaks must be separated by a time interval theta or have the intermediate flow drop below 75% of the minimum of both peaks.

   $$\Theta > 5\ days + log(A) \quad or \quad Q_{min} < 75\%\,[Q_1, Q_2] \tag{1}$$

   where A is the catchment area in square miles. In the revised manuscript, we will elaborate on these details in Section 2.3.

2. Flood magnitudes are displayed by ranks. This requires better introduction, explanation and guidance for the reader when explaining the results.

   Thank you for bringing our attention to this. Indeed, the explanation of the ranks was rather brief in the original manuscript. In the revised manuscript, we will clarify the usage of the ranks and provide a comprehensive explanation for better interpretation of the results.

3. The authors could elaborate more about the uncertainties of their simulations and projections. In particular, the role of the deep learning approach on the overall uncertainty could be discussed in some more detail.

   Thank you for pointing this out. We will add a dedicated section to discuss the uncertainty regarding LSTM due to randomness and the selection of hyperparameters.

Specific comments

1. The results presented in the abstract could contain more specific, quantitative information. The author may also consider adding details on the data and methods used (which deep learning model used etc.)

   Thank you for the suggestion. We will add more details to the abstract.

2. The introduction lacks a clears research question or research objective. Furthermore, the introduction should also introduce and explain trans-basin floods in some more detail. I also suggest to highlighting the relevance of rain-on-snow events compared to other flood types in Germany in some more detail.

   Thank you for the suggestion. We will revise the introduction and add more details on trans-basin floods and rain-on-snow floods.

3. L39 add direction of change

   We will revise this line.

4. L54 why?

   A warming of up to 4°C increases the possibility of liquid precipitation, while there is still enough snow pack left. If temperature increases more, there will be far more days without snow covered terrain and thus the risk of ROS decreases. We will clarify this in the revised manuscript.

5. L55-56 this sentence is not complete

   Thank you. We will revise the sentence.

6. L70-78 on which spatial scales? What about uncertainties in the (hydrological) impact models?

   We will revise this part and add the information about the hydrological model and the spatial scale of the referenced study from Sezen et al. (2020). They use the lumped conceptual hydrological model GR6J to simulate streamflow for three small catchments ($60 - 160\ km^2$) located in Slovenia.

7. L93 already mentioned

   Thank you for pointing this out. We will remove the repeated mentioning of the GRDC dataset.

8. L96 add spatial scale of ERA data

   We will add the spatial scale of ERA5.

9. L107 would it make sense to apply an ensemble of LSTMs to learn more about the uncertainty of the deep learning approach?

   Thank you for your commend. Indeed, uncertainty of deep learning models stemming from randomness in the training processes (e.g., the random initialization of weights) can be substantial. However, we show in Figure 2 that the effect of the segment of the training data that was used for training is much stronger than the effect of initial randomness. We will clarify this point in the revised manuscript.

10. L135 check for consistent use of IGsm and IGsnw

    Thank you for pointing this out. We will correct the IG_{snm} to IG_{sm} and IG_{sm} to IG_{snw}.

11. L186 the authors should provide more information on how they performed the bias correction.

    We will add additional details about the bias correction.

12. L187 This could need a clearer explanation. Maybe highlighting the difference between ERA5 and ERA-Interim could help.

    Thank you for your suggestion. We will introduce the data and the idea of the usage of the downscaled ERA-Interim simulations more clearly.

13. L230-235 What about anthropogenic activities in the study catchments in general? I would guess that this is an issue. How does this impact the results?

    Thank you for bringing this important point up. Anthropogenic activities likely will impact the other gauges, too. The impact, however, may vary considerably depending on the type of activities and their persistency. If the activities happen on a regular basis over the entire period the LSTM should be able to learn the anthropogenic impact similar to learning a systematic bias within a climate model. If the activities do not happen on a regular basis or are inconsistent within the training data (e.g. structures built during the time period) the model is not able to learn the impact. This is likely the reason why some gauges clearly underperform compared to other gauges with a similar catchment area. In single cases, the lower performance appears to have an impact on the peak flow and the peak timing, but there is no clear connection between the NSE from Figure 2 and the strong dips in Figure 4. It is worth mentioning, that this should not have an impact of the flood generating processes at any given flood peak. Yet, if all peaks of specific flood generating processes are systematically underestimated, it might result in systematically poorer performance for a given process. Given the good agreement to Uhlemann et al. (2010) regarding winter widespread floods and the high number at ROS floods even at gauges with a comparably poorer performance, it is unlikely the case for the ROS process. We will clarify this

in the revised manuscript.

14. Figure 3: I do not understand the orange line in combination with lines 240-244.

    The orange line is meant to indicate the fraction of simulated peaks that have a good timing with a maximum shift by 1 day compared to observations. E.g. if the orange line indicates that 80% of peaks have a timing of +-1 day, the other 20% have a timing of +-2 or +-3 days. We will clarify this description in the revised manuscript.

15. Figure 4: The LSTM predicts floods not contained in the observation data. What does this imply for future projections?

    Thank you for your comment. If the LSTM predicts floods that are not contained in the observation data there was a signal (e.g., a rainfall or snowmelt peak) in the forcing data (i.e., ERA5). There are several possible explanations for the occurrence of those floods. Due to the spatial resolution of ERA5, the actual rainfall event could have happened outside of the catchment area (which would be relevant especially for smaller catchments where the spatial resolution of ERA5 has a higher impact) as ERA5 is not able to better resolve the event. Given that winter rainfall events are generally more widespread this would likely rather impact more localized summer events. Additionally mentioned flood events occur in all catchments independent of their size. Another explanation is that rainfall peaks are over- or underestimated in ERA5 leading to over- and underestimated floods predicted by the LSTM. If we assume that ERA5 is consistent and the distribution of rainfall events is realistic in a climatological sense, this should not affect the results of the future projections. Also, the when considering widespread floods, the events detected in the LSTM simulations are comparable to those of Uhlemann et al. (2010). We will clarify this in the revised manuscript.

16. L290-295 I think these data sets need a better introduction in chapter 2.

    We will add more detail to the description of the CORDEX data.

17. Figure 7c, f (plus corresponding text, L293): I do not see this improvement.

    Thank you for the comment. The lines 292-294 could be clearer about the message of the figure. If the ERA5-trained LSTMs are driven with the bias corrected data from the CORDEX evaluation runs the simulation is much better in terms of NSE (Figure 7c) and produces a less noisy streamflow (Figure 7f) compared to driving the ERA5-trained LSTMs with the raw CORDEX evaluation runs (Figure 7a and 7d). This is not surprising, as the LSTMs that learned the biases present in ERA5 and are driven by the raw data cannot handle the biases coming from the CORDEX RCMs, thus bias-correcting the data towards ERA5 makes it easier for the LSTM to handle the CORDEX data as input. Given that for each of the RCMs an evaluation run with downscaled ERA-Interim data was available, we trained the LSTM directly with the evaluation runs to let the model learn the biases of the RCMs directly and to skip the bias-correction if possible. However, this approach (Figure 7b) shows a similar performance to the approach with the ERA5-trained LSTM and bias-corrected data (Figure 7c), although the streamflow in Figure 7e appears to have less spikes than in Figure 7f. Additionally, the approach to train the model directly with the RCM evaluation runs comes with some drawbacks such as the much higher number of LSTM models that have to be trained and a much shorter available time period for training. We will adjust the lines 292-294 to be clearer about this improvement and our overall idea behind the figure.

18. With reference to the skill in predicting high-flows, this should read a bit more critical. How well does the deep learning approach compares against a regular hydrological model? Is it possible to assess/discuss this?

    Thank you for this comment. We will add a part discussing the differences in performance between LSTM and hydrological models. Recent studies show that dep learning models perform as well or even better than regular hydrological models (Nearing et al., 2024; Frame et al., 2022). Those studies show that deep learning models do not have a general problem predicting high flows when there are comparable events within the training data. However, even when such events are not included in the training data, deep learning models can perform better than regular hydrological models (Frame et al., 2022).

19. L344 add a reference to Figure 10 underlining the spatial information

Thank you. We will add a reference to Figure 10 in the revised manuscript.

20. Figure 10: is it possible to indicate the significance of changes here?

    Thank you for your suggestion. We will modify the figure to show the significance of the changes.

21. Figure 11: notches in the boxplot may indicate the significance of changes; Fig. 11 c) why "all" smaller than individual river basins

    Indeed, as the box plot indicates there might be significant changes in the characteristics of ROS floods in some of the study basins.
    "All" is smaller in this case, simply because the total area is larger. Additionally, not only ROS floods contribute to the trans-basin floods, but also snowmelt and rainfall driven floods. If there was e.g. a large meteorological event that leads to ROS events in the Danube basin, but to rainfall driven floods in the Rhine and Weser basins, the absolute ROS affected area is the same for the "all" and "Danube" category, but the total catchment area in "Danube" is smaller and thus has a larger fractional ROS affected area. We will clarify better the meaning of the label "all" in the revised caption.

22. Figure 12 (c): Do I understand correctly that ROS events are currently by far the most dominant flood types in Germany? I can hardly believe that.

    ROS floods are an important driver of floods in Germany, but not the dominant one. According to Tarasova et al. (2023), the share of annual maximum floods that are driven by ROS is between 20 and 30% in individual German catchments. The share of winter trans-basin floods that are likely driven by ROS is even larger (53 out of 80 documented events, Uhlemann et al. (2010)). Therefore, the proportion of trans-basin floods attributed to ROS in our study is in line with previous reports. We will clarify this point in the revised manuscript referring to the previous reports of trans-basin ROS events.

23. Figure 13: maybe the small statement in the text on changes in the seasonality is enough and does not require this figure which does not provide so many extra information.

    Thank you for your suggestion. Indeed, this information can be presented in the text. We will follow reviewer's suggestion and remove this figure.

24. L449 I wonder if percentage change is an appropriate measure here. If we assume the same absolute decrease in the number of days with snow cover in lowlands and the Alps, then the percentage decrease is of course smaller in the Alps since the total number of snow days is generally larger high-mountain regions.

    Thank you for your comment. We understand reviewer's concern. However, we think that it makes sense in the context that this impacts the lowlands much more than the Alps in regard to the decrease in ROS frequency because with the reduction of the same absolute numbers of snow days the potential for ROS floods is reduced a lot more in the lowlands than in the Alps as there are still a lot more days with snow cover left. Therefore, we plan to keep percentage change as the metric.

25. L486-490: repetition, this has already been mentioned in the method section

    Thank you. We will revise this part.

References

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S.: Deep learning rainfall–runoff predictions of extreme events, Hydrology and Earth System Sciences, 26, 3377–3392, https://doi.org/10.5194/hess-26-3377-2022, 2022.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., et al.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, 2024.

Sezen, C., Šraj, M., Medved, A., and Bezak, N.: Investigation of Rain-On-Snow Floods under Climate Change, Applied Sciences, 10, https://doi.org/10.3390/app10041242, 2020.

Tarasova, L., Lun, D., Merz, R., Blöschl, G., Basso, S., Bertola, M., Miniussi, A., Rakovec, O., Samaniego, L., Thober, S., and Kumar, R.: Shifts in flood generation processes exacerbate regional flood anomalies in Europe, Communications Earth and Environment, 4, https://doi.org/10.1038/s43247-023-00714-8, 2023.

Uhlemann, S., Thieken, A. H., and Merz, B.: A consistent set of trans-basin floods in Germany between 1952-2002, Hydrology and Earth System Sciences, 14, 1277–1295, https://doi.org/10.5194/hess-14-1277-2010, 2010.

USWRC: Guidelines for determining flood flow frequency, 17, US Water Resources Council, Hydrology Committee, 1975.