

Dear Editors and Reviewers,

Thank you for taking the time to provide constructive and thorough feedback that has helped improve our manuscript. Hereby, we would like to provide our main changes and point-by-point reply to the first and second reviewers' comments, abbreviated RC1 and RC2, respectively. Finally, we respond to the remark of the editor. Our replies are marked by 'Authors', and reference to where changes have been made in the track-changes version of the manuscript are marked by 'Change'.

First, we would like to outline the major changes performed:

- A. *As a response to RC1's first general comment (number 1.1 below), we added results of false alarm rate and probability of detection to our analysis (Fig. 7 in the revised manuscript).*
- B. *To address RC1's third general comment (number 1.3 below), we wanted to recalibrate HBV using NSE as a loss function to make sure LSTM does not have any advantages due to the difference in applied loss function. After this manuscript was first submitted, the operational HBV model set-up and calibration period has been changed. The current operational HBV model is calibrated on daily discharge divided at midnight (equal to what is used for LSTM) over the period 1.9.2009-31.8.2024. Because the previous set-up lacks documentation (e.g. on applied loss function and daily divide), we chose to update our analyses to match the current calibration period and set-up of the operation HBV. Thus, we needed to retrain LSTM, using the same cross-validation procedure as presented in the preprint, but for a different seen period to match that of HBV. Further, all evaluations are performed over the new unseen period 1.9.1994-31.8.2009. This change affects all result figures. The results are similar, indicating robustness of the conclusions. The most notable changes from the results presented in the preprint are an increase in too-early simulated flood peaks by HBV, and more catchments having notable difference in runoff coefficient between the calibration and evaluation period.*
- C. *As a response to RC2's first general comment (number 2.1 below), we assess the sensitivity of chosen flood classification thresholds by reproducing figure 5 (flood peak timing) and 6 (percent error in flood peak magnitude) using two threshold alternatives, both allowing for more mixed events. Results are provided in appendix together with a histogram over all flood events' fractional rainfall contribution (Fig. C3-C7) and are commented on in the revised manuscript.*
- D. *Following the advice of RC2 (number 2.4 below), we moved part of the calibration/training set-up of the models to appendix (Appendix A for LSTM and Appendix B for HBV in the revised manuscript). Further, LSTM is introduced before HBV in the Data and methods chapter.*

In the following, point-by-point replies and changes following each reviewer comment are provided.

First reviewer: general comments

- 1.1 RC1 The authors state that the main objective of this study is to evaluate the potential of LSTM for operational use in flood simulations and flood forecasting in snow-influenced regions. However, the model is only tested against observed (i.e., really occurred) peak flow events. How does the LSTM perform compared to the HBV model in terms of simulating false positive events (i.e., floods that have not been observed but simulated)? I think this is another important aspect which determines the model's suitability for different purposes such as flood warning. A confusion matrix comparing both models could provide some insights in this regard.

Authors	We thank the reviewer for this valuable comment, and agree that an assessment of false positives is an important aspect when considering a model for operational use. To align with our study design, we add such an analysis performed separately for the different flood types. In this regard, a confusion matrix is not optimal, as defining flood generating process (FGP) of non-flood events does not make sense. A confusion matrix based on timeseries data also suffers from dependency of datapoints, and relatively small mismatches in time may have a notable impact on the results. Rather, we have chosen to compute the related summary metrics false alarm rate (FAR) and probability of detection (POD) which assess to what degree a model produces false positives and true positives, respectively. FAR and POD are based on independent flood events, and fit well with our study design as it can straightforwardly be performed for different flood types. Further, simulated and observed events can be matched using both the pinned and floating methods. The method is introduced in Sect. 2.7 and results are shown in Fig. 7 in the revised manuscript. Related edits are made in the text.
Change	L16-17 (Abstract), L306-317 (Data and methods), L371-378 and Fig. 7 (Results), L450-451 (Discussion), L522-523 (Conclusions)
1.2	RC1
	The advantage of a process-orientated hydrological model compared to LSTM is that the flood generating processes can be directly derived from the model output using the described approach since HBV also simulates snowmelt. LSTM and other deep learning approaches are black-box models and do not support this process mapping. There are a few approaches that attempt to combine LSTM with explainable AI (e.g. De la Fuente et al., 2024; Czakay et al., 2025 and references within) to better capture the hydrometeorological process relationships. I think that AI approaches for hydrological applications will evolve in this direction. So, if not by extending this study in this direction, this topic should at least be addressed in the discussion.
Authors	We agree that this is an interesting topic. As suggested by the reviewer, we have addressed this topic (and the related comment 1.8) in the discussion.
Change	L480-493
1.3	RC1
	It seems that the HBV model and the LSTM have been calibrated/trained by minimizing different objective functions (sum of squared errors + squared total volume errors vs. NSE). This may impact the relative ability of both approaches to reproduce peak flow events. For the comparison of the LSTM against the benchmark HBV model this could be a serious issue. It needs to become clear that LSTM does not have an advantage in modelling flood events due to a better and more specific training for simulating high discharge compared to the benchmark model.
Authors	We agree that the different loss function can have an impact on the results. As already described in bullet point B of our list of major changes provided above, we have now recalibrated the operational HBV model using NSE as a loss function. Because the set-up and calibration period of the operational HBV model has been changed since we first submitted this manuscript, and the former set-up lacks documentation, we chose to use the most up-to-date model set-up and calibration period. Thus, LSTM has also been retrained, using the exact same procedure as presented in the preprint, but using the new calibration period 1.9.2009-31.8.2024. One of the hyperparameters, i.e. batch size, had a change in optimal value. We note that the left-out-fold results were not notably sensitive to this

hyperparameter. All evaluations are performed over the new unseen period 1.9.1994-31.8.2009. The new results resemble the ones presented in the preprint, indicating robustness of the conclusions. As mentioned above, the most notable changes from the results presented in the preprint are an increase in too-early simulated flood peaks by HBV, and more catchments having a notable difference in runoff coefficient between the calibration and evaluation period. The latter is expected due to temporal changes in the measurement network being more notable when the calibration period no longer is 'in the middle of' the evaluation period.

Change All result figures are updated with the new results. Reported resulting numbers and text are updated accordingly throughout the manuscript. Changes in the Methods chapter are made to match the new periods and HBV modelling.

First reviewer: specific comments

- | | | |
|-----|---------|--|
| 1.4 | RC1 | Introduction or Data: Since the study focuses on Norway, the relevance of snowmelt, rainfall and a combination of both for flood generation should be better demonstrated for this region. |
| | Authors | We added a new section in the Data and methods chapter introducing the hydroclimatology and its relevance for floods in the study region (Sect. 2.1 in revised manuscript). |
| | Change | L73-86 |
| 1.5 | | |
| | RC1 | Introduction: LSTM is one deep learning approach among others. The literature review should be a bit broader in this regard. Since LSTMs often perform better than other methods, this actually provides a good reason for why LSTM is applied in this study. |
| | Authors | We added the following sentence addressing this: <i>“Further, rainfall-runoff modelling by LSTM surpasses other deep learning techniques, such as traditional recurrent neural networks (Kratzert et al., 2018), feed-forward neural networks (Hagen et al., 2023), and Transformer-based models (Liu et al., 2025).”</i> |
| | Change | L29-31 |
| 1.6 | | |
| | RC1 | L42 I am at least aware of the beforementioned study that apply a LSTM to simulate rain-on-snow floods in Germany |
| | Authors | Thank you for the suggestion. In the study you refer to, Czakay et al. (2025) indeed apply LSTM to simulate rain-on-snow floods in Germany. However, their evaluation of LSTM’s ability to simulate floods (Fig. 3-5) are performed over all identified flood events, and not rain-on-snow events specifically. Thus, our sentence hold. |
| | Change | No change in revised manuscript. |
| 1.7 | | |
| | RC1 | L87 Please indicate percentage or number of catchments being partly glacier cover. |
| | Authors | We added the following sentence: <i>“Specifically, 21 (12) of the 200 catchments have more than 3 % (10 %) of their catchment area covered by glaciers in 2019.”</i> |
| | Change | L109-110 |
| 1.8 | | |
| | RC1 | L92 Comparability of the SeNorge snow model and the output of the HBV snow routine: Here, I was wondering if it would make sense to evaluate the ability of the benchmark model to correctly model the specific flood type as classified by the approach in section 2.6 using the |

		data described in 2.2 and the output of the catchment specific HBV. On the other hand, this cannot be tested for the LSTM. Same direction as my GC (2).
	Authors	We agree that this is an interesting topic. As suggested by the reviewer in his general comment 1.2, we have addressed the topic in the discussion.
	Change	L480-493
1.9	RC1	Table 1 Just for curiosity: has it been tested how the LSTM performs when using only subsets of these catchment attributes? Alternatively, could specific attributes be identified as important predictors? Sometimes this leads to some hardly explainable surprises.
	Authors	We find this a very interesting topic. Although outside the scope of the current study, we plan to do so in the future (e.g. similar to Sourya et al., 2025; doi: 10.1016/j.jhydrol.2025.134270) to further improve and understand the model for our region.
	Change	No change in the revised manuscript.
1.10	RC1	L126 Was the same objective function used to optimize the LSTM? If not, this could have an impact on the ability of simulating peak flows.
	Authors	In the revised manuscript, we have changed HBV's loss function to NSE and kept catchment-averaged NSE for LSTM. For details, please see bullet point B on the first page, and our answer to comment number 1.3.
	Change	Ref. comment number 1.3. Specifically, we state the applied loss function for LSTM on L195-196 and HBV on L235-236
1.11	RC1	Section 2.4 Is the LSTM applied on a gridded basis or for catchment averages?
	Authors	The LSTM is applied for catchment averages. We already specify that the meteorological input time series are catchment averages (L121-122 and L224-226 in the revised manuscript, no change from the preprint). Further, we have now clarified that LSTM is employed as a lumped rainfall-runoff model.
	Change	L190
1.12	RC1	Sections 2.4 and 2.5 I suggest moving the last paragraph of 2.4 to 2.5 and the first paragraph of 2.5 to 2.4.
	Authors	Thank you for this suggestion, we agree that this is a logical change. We have now made changes to the sections on LSTM in the data and methods chapter, following both this comment and comment 2.4. In the revised manuscript, there is one section (Sect. 2.4) about LSTM in the data and methods chapter, and further details on the training of LSTM in Appendix A.
	Change	L182-224 and L542-574
1.13	RC1	Section 2.6 The approach of detecting flood generating processes is comparatively simple but effective. Still, I suggest putting this into context of other more sophisticated approaches of flood type differentiation.
	Authors	We have added the following text in Section 2.6: " <i>We selected this approach from the range of available flood classification schemes (Tarasova et al., 2019) because it provides a simple yet effective distinction between rainfall and snowmelt generated floods and has been successfully applied in studies conducted within our region (Yan et al., 2023; Vormoor et al., 2016, 2015).</i> " We also address the topic in a new paragraph in the Discussion chapter (related to comment 1.2).

Change	L254-256 and 480-493	
1.14	RC1	L269 I am a bit surprised about the low number of mixed floods. Were the same thresholds regarding snowmelt- and rainfall contribution used as in the cited reference? Generally, it is worth adding one or two more sentences on this procedure in section 2.6 so that the reader does not need to search for the reference.
	Authors	The thresholds are the same as used in the cited reference. We have now clarified this adding the following “ <i>Using the thresholds proposed by Vormoor et al. (2015) ...</i> ” Further, we added a figure in appendix (Fig. C3) showing the distribution of fractional rainfall to all flood events, and two optional classification threshold pairs for producing timing and PE plots (ref. comment number 2.1). Here one can see the relatively low number of mixed flood events as compared to snowmelt and rainfall generated flood event. The results using the alternative thresholds (which allows for more mixed events) are very similar to our main results (Fig. C4-C7 as compared to Fig. 5-6).
Change	L260, L262-365 and Fig. C3-C7	
1.15	RC1	L281 For hydrological models, the goodness-of-fit is often higher due to low model errors during winter low flows, particularly in snow-dominated catchments. Is this also the case for LSTM? According to Figure 4 (c), probably not as pronounced as for HBV.
	Authors	We have extended the comment we had about this in the Discussion, to the following: “ <i>Many of the snowmelt dominated catchments were among the catchments with the highest NSE scores for HBV, in line with previous studies demonstrating that models typically achieve higher NSE and KGE values in catchments with strong seasonality (Ruzzante et al., 2026; Anderson and Radic, 2022). However, for LSTM in terms of NSE, and for both models in terms of KGE, there were no distinct relation between catchment dominant FGP and model performance score.</i> ”
Change	L422-426	
1.16	RC1	Figure 5 should indicate the number of events per class either in the bar labels or in the figure caption.
	Authors	Good idea, we now include the number of events per class in Fig. 5.
Change	Fig. 5 and an extra sentence in its caption.	
1.17	RC1	Figure 10. Station numbers in the x-axis might be confusing. I suggest numbering from 1-103, or indicating that numbers are station numbers or catchment IDs, respectively.
	Authors	Good point, we have now clarified by adding the x-axis label “ <i>Catchment ID</i> ”
Change	Fig. 11	
1.18	RC1	L389 This sentence needs reformulation. What aspects? Maybe “The relevance of ... differ between”. In addition, these “aspects” may also differ for different spatial and temporal scales, i.e., flood type dominance, relevance of catchment (storage) characteristics, antecedent conditions, flood durations etc.
	Authors	We agree that the sentence is vague, and during revision of our manuscript, we do not see that it adds much value to the discussion. We have therefore deleted it in the revised manuscript.
Change	L478-479	

First reviewer: Technical corrections

- 1.19 RC1 L2 I suggest writing "... most successful deep learning models..."
Authors We wish to keep our original formulation, as LSTM is the most successful over large regions also when compared to process-based models (demonstrated by several large-scale studies, e.g. references in first paragraph of the Introduction).
Change No change in manuscript
- 1.20 RC1 L114 Abbreviation NVE has already been introduced
Authors Fixed.
Change L233
- 1.21 RC1 Check the use of "percent" vs. "percentage". For example, I think it is called mean absolute percentage error.
Authors We agree that 'mean absolute percentage error' is more commonly used than 'mean absolute percent error', and have made changes in the manuscript accordingly.
Change Table 2 and its caption, L318, L321, captions of Fig. 8-9, L450 and L528
- 1.22 RC1 When referred to a figure in the main text, it is "Figure #" not "Fig. #"; "(Fig. #)" is correct.
Authors We wish to follow the HESS guidelines which states "*The abbreviation "Fig." should be used when it appears in running text and should be followed by a number unless it comes at the beginning of a sentence, e.g.: "The results are depicted in Fig. 5. Figure 9 reveals that..."*"
<https://www.hydrology-and-earth-system-sciences.net/submission.html#figurestable>
Change No change in revised manuscript.
- 1.23 RC1 L351 "has previously shown"
Authors Fixed.
Change L432
- 1.24 RC1 L365 I suggest shortening this to "...for different flood types..."
Authors Fixed.
Change L453
- 1.25 RC1 L420 "errors were notably smaller" rather than "...better"
Authors Fixed.
Change L525
- 1.26 RC1 L511 Reference Langsholt is missing an „E.“
Authors Fixed.
Change L677

Second reviewer: general comments

- 2.1 RC2 The authors assess the performance of the LSTM for snowmelt, rainfall, and mixed events separately. Although the definition of the different event types is briefly described in the

methods section 2.6, I would recommend to expand on the procedure of defining the flood generation processes. I assume that boundaries between the different processes are generally difficult to define, and it would be interesting to see how a change in the distribution of processes (e.g., more mixed events, less rainfall events, ...) may influence the predictions obtained with the LSTM in each group, e.g., when looking at Figure 5 (or if this has an impact at all?). I do not expect the authors to add a detailed sensitivity analysis, but recommend adding this to the discussion section.

Authors Thank you for this good suggestion. To assess this, we produced Fig. 5 (flood peak timing) and Fig. 6 (Percent error of flood peak magnitude) using two alternatives to our original thresholds of 1/3 and 2/3. One symmetric alternative (0.2 and 0.8) and one asymmetric alternative (0.3 and 0.9), increasing the number of mixed events from 387 to 859 and 1073, respectively. Both alternatives produced very similar results to the original ones. The results using the two alternatives are provided in the Appendix (Fig. C4-C7). We included a Figure (i.e. Fig. C3) showing the histogram of fractional rainfall contribution of all flood events, with the inserted thresholds, to clarify the distribution of flood events (and thereby also addressing comment number 1.14).

Change Fig. C3-C7, L262-265, L361-362, L364-365 and L445-446

2.2 RC2 In my view, the discussion is quite methods-oriented, which is well reasoned in itself. However, I am missing potential (physical) explanations on why both models performed differently for the different event types, for example, potential reasons for the differences in timing errors between snowmelt and rainfall generated events? I understand that interpreting the results obtained with the LSTM can be more challenging compared to the HBV model. Nevertheless, I would encourage the authors to include potential explanations into the discussion section, including suitable references of previous studies.

Authors In Sect. 4.2 in our preprint, we included a potential explanation for the differences in timing errors between snowmelt and rainfall generated events: *“As snowmelt generated floods react to snowmelt typically spanning days to weeks, the floods often last multiple days and may not have a very distinct peak as compared to rainfall generated floods. Relatively small differences between peak discharge and discharge magnitudes in adjacent days may explain the lower percentage of correctly simulated peak days for snowmelt generated floods.”* We suspect this to be the main reason for the detected timing difference. We added a reference to a figure from another paper which exemplifies this. Further, to also address comment 2.8, we added that the same reasoning may explain why results for mixed events were better than snowmelt generated events but worse than rainfall generated events. After a new literature search (ref. comment number 2.9), we still did not find other studies than (the already mentioned) Brunner et al. (2021) that report flood-type dependent flood peak performance results.

Change Edits in Sect. 4.2 (L443-459)

Second reviewer: minor comments

2.3 RC2 L 42-45: This sounds as if you trained the LSTM separately for snowmelt and rainfall events, but as far as I understood, you trained it over all catchments and evaluated its predictions separately. Please clarify in the text.

	Authors	Thank you for pointing out this unclarity. Your understanding is correct, and we have edited the sentence to clarify what we mean: <i>“Nevertheless, to our knowledge, existing evaluations of LSTM’s ability to simulate floods are performed over all identified flood events, without evaluating events generated by different processes separately.”</i>
	Change	L49
2.4	RC2	L 65: In my opinion, the data and method section is excessively long. The authors may want to consider creating a Supporting Information document and moving parts of this section to that document, e.g., model details in sections 2.3, 2.4, 2.5.
	Authors	We agree that this a good idea and have moved details concerning LSTM training and HBV calibration to the Appendix A and Appendix B, respectively. We now also present LSTM before HBV in the Data and methods chapter.
	Change	L132-164, L182-224, L230-240, L542-591
2.5	RC2	L 117: The 103 stations have been repeated many times; consider consolidating these parts.
	Authors	We agree and have reduced the number of repetitions.
	Change	L238-240, L267 and L330.
2.6	RC2	L 196-199: This paragraph needs references.
	Authors	References added to the first sentence.
	Change	L547
2.7	RC2	L 227-231: I do not think that it is necessary to mention the periods before 1994 here, as 15 years can already be considered a sufficient length.
	Authors	We agree and have removed the mentioning of older periods.
	Change	L268-272
2.8	RC2	Figure 5: Why do the authors think the peak timing of snowmelt events is harder to predict than mixed events? Intuitively, one could expect that it would be the other way around? This may be added to the discussion (see also general comment).
	Authors	We suspect it can be explained by the same reasoning as provided for the difference between snowmelt and rainfall generated flood events, i.e. that it is generally harder to predict the correct flood peak day of longer-lasting events with less distinct peaks (ref. comment number 2.2). We have added the following in Sect. 4.2: <i>“The same reasoning may explain the intermediate position of the timing results for mixed flood (better than snowmelt generated floods but worse than rainfall generated floods).”</i>
	Change	L451-452
2.9	RC2	L 368: Any other references here?
	Authors	As also mentioned in our answer to comment 2.2, we did a new literature search, without finding additional studies that report flood-type dependent flood peak performance results.
	Change	No change in the revised manuscript.
2.10	RC2	L 390: This sentence needs a reference.
	Authors	The sentence is deleted in the revised manuscript (ref. comment number 1.18).
	Change	L478-479

Remark by editor

- Editor The ms has been well received and I agree it is well written and structured. I align with the referees that some more in-depth comparison between LSTM and HBV in terms of false positives is an important improvement. I look forward having the extra analysis included and discussed in the revised ms.
- Authors We thank the editor for taking over the handling of our manuscript, as well as the positive feedback. The comparison between HBV and LSTM in terms of false positives for different types of floods is now included as Fig. 7 in the revised manuscript.
- Change We refer to our response and changes provided to comment 1.1.