Overview of changes in the revised manuscript

We thank the reviewers for their valuable comments and suggestions. To address these comments, we implemented the following changes in the revised manuscript.

Major changes:

- To address Reviewer 1's comments and the community comment regarding the model's sensitivity to climatic changes, we have:
    - Conducted and included a complementary analysis evaluating the model's generalizability with respect to trend preservation (**new Section 3.4)**
    - Added a new figure illustrating these results (**Figure 6** in the revised manuscript)
- The **abstract** has been revised to include clearer and more quantitative results, thereby explicitly demonstrating the added value of the hybrid model. The abstract now also incorporates key findings from the new generalizability analysis, further supporting the benefits of the hybrid approach.
- In response to multiple comments from both reviewers concerning the clarity of the Methods section, we improved the consistency, clarity, and logical flow of the Methods section by:
    - Clarifying the distinction between the in-sample and regional HBV model setups (S**ection 2.2.1)**
    - Adding a schematic overview of the evaluation framework as a new subpanel (**Figure 2a)**
    - Revising and expanding the description of the evaluation framework to improve readability and clarity (S**ection 2.4)**
- The discussion section has been extended (**Section 4.1)** by:
    - Addressing the influence of the chosen objective function on the representation of low-flow performance
    - Discussing model generalizability in the context of changing climate conditions

Minor changes:

- Key results from the newly added analysis have been incorporated into the conclusions (**Section 5**).
- The description and referencing of the hybrid model have been harmonized and refined throughout the manuscript.
- The **Code and data availability statement** has been expanded to explicitly acknowledge and credit the original developers of the hybrid model
- Terminology has been standardized and used consistently across all sections
- Figure style and visualization have been made consistent throughout the manuscript.
- Minor editorial changes have been made to improve readability and overall narrative flow.

In the following we provide the original reviewer comments in **black,** our initial response to these comments in **blue**, and the final changes implemented in the revised manuscript are highlighted in **plum**.

Reviewer 1

The manuscript evaluates the ability of different model types (HBV, LSTM, and a hybrid model) to predict river streamflow under different climate conditions, particularly when the training/calibration period differs from the testing/validation/prediction period. This issue is critical when applying machine learning models to future climate-change impact studies. The manuscript is well written, the experimental design is appropriate for the scientific questions, and the results are clearly illustrated. I have several major comments that I would like to discuss with the authors. If these can be addressed, I would recommend the paper for publication.

Thank you very much for your positive feedback and your constructive and valuable comments. Please find our responses to your comments in blue following your comments in black.

First, I think a sensitivity test should be conducted. Before applying the models to the warmer period, perturb the input variables (such as temperature or precipitation) and evaluate how the models respond to these changes. This is relevent for the following analysis, maybe different model is sensitive, others are not.

Thank you for your thoughtful comment. The question you raise is indeed central to what we aim to evaluate in this study. In our definition of generalizability, we require that model performance remains stable across different time periods. If an individual model shows a substantial increase or decrease in performance between periods, this suggests that the model may be sensitive to an external driver—such as temperature or precipitation—and therefore not fully robust under changing climate conditions.

We chose to split our analysis according to periods with a dominant temperature change because (a) temperature exhibits a pronounced shift between the two periods, and (b) this shift is associated with a clear reduction in snow fraction. This trend is expected to intensify with continued warming and thus represents a meaningful challenge for hydrological model setups.

We appreciate your point that one could first examine intrinsic model sensitivity. However, if any model were strongly sensitive/insensitive to temperature, this would manifest itself in our evaluation as well. Martel et al. (2025) conducted such a sensitivity analysis for an LSTM and several conceptual models, showing that the LSTM exhibited slightly higher sensitivity to temperature changes but slightly lower sensitivity to precipitation changes compared to the conceptual models. It is worth noting, though, that their experiments used substantially larger imposed temperature changes (+3°C and +6°C) than the observed trends in our study (approx. +1 °C). Their focus was on future climate projections, which is highly relevant but somewhat different from the scope of our work. Nevertheless, such a sensitivity analysis could be a natural extension of our current study.

Martel et al 2025: https://hess.copernicus.org/articles/29/2811/2025/

Following your suggestion and a community comment, we would like to propose a complementary perspective on sensitivity that we believe aligns even better with the aims of the paper. Specifically, we quantify the modelled changes in key streamflow metrics between the warm (warmest) and the cold period and comparing these changes to the observed ones. This allows us to assess not only whether model performance changes across periods (i.e. current analysis), but also whether each model appropriately captures the magnitude and direction of change in the respective metric in response to mainly changes in temperature (i.e. new

analysis). That is, we test whether the different models can represent the sensitivity of streamflow to changes in temperature.

We conducted and included a complementary analysis evaluating the model's generalizability with respect to trend preservation, that is the model's sensitivity to changes.

We added the new section "3.4 Model generalizability in preserving trends" to the manuscript

- P17f, L377-423: "Next, we look at generalizability from the perspective of trend preservation. [...] The LSTM generally shows the same direction of change as observations, but has a slightly enhanced sensitivity compared to observations and the other models."
- We added a new figure illustrating these results (see Figure 6 in the revised manuscript and below).

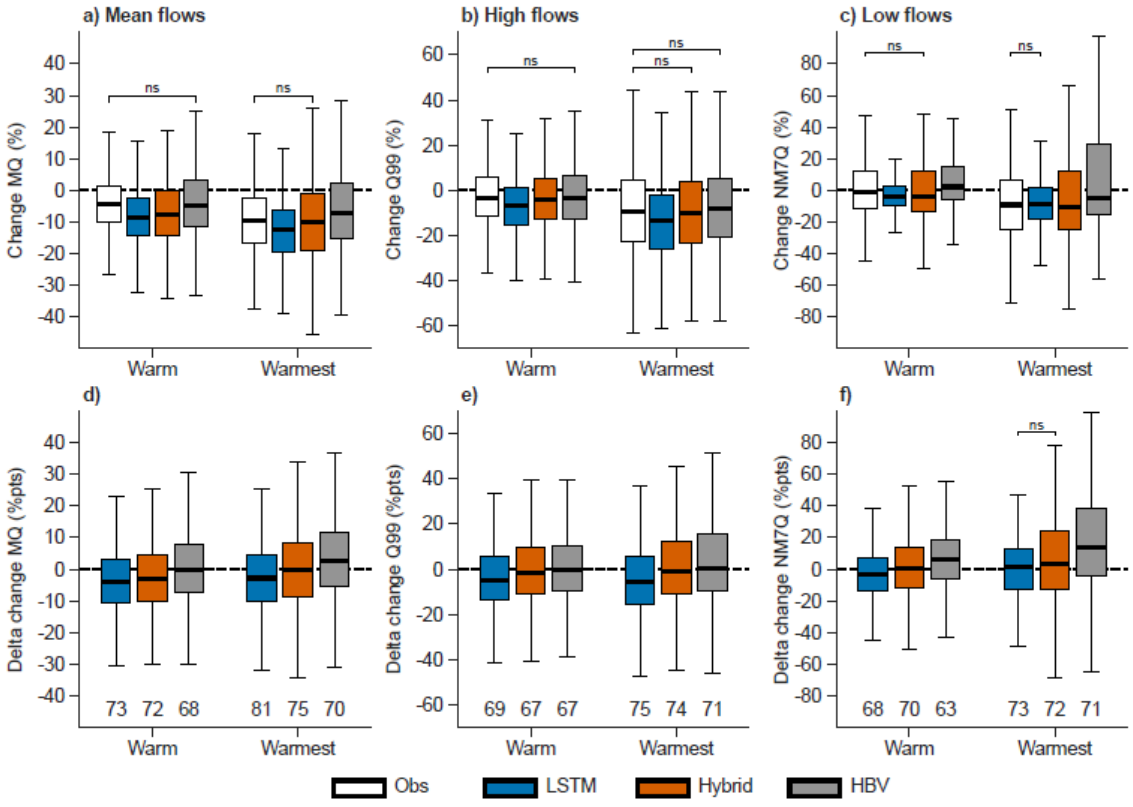

Figure 6: Comparison of observed and simulated streamflow trends between the different periods in all catchments. (a-c) Relative differences in (a) mean flow (MQ; mean annual streamflow), (b) high flows (Q99; mean 99th percentile of daily streamflow in a year), and (c) low flows (NM7Q; mean annual NM7Q) between the warm/warmest and cold period. (d-f) Catchment-wise differences in the (d) mean flow, (e) high flow and (f) low flow signals between the model and observations. Above the boxplots, we show the results of a two-sided sign test. The model pairs where the p-value is above 0.05 are highlighted by 'ns'. For all other pairs, significant differences are found. Below the boxplots in (d-f), we show the model agreement with observations, that is the percentage of catchments with an agreement on the sign of change.

We added a short section in the discussion on the connection of generalizability and meteorological forcings in the discussion (Section 4.1)

P21, L479-84: “The lower generalizability of the LSTM compared to the hybrid and HBV model also holds when comparing observed and simulated streamflow signals. The LSTM simulations show a higher sensitivity to changes in meteorological forcings than observations (Figure 6), that is stronger declines in mean and high flows. This might indicate that when moving to even warmer climate conditions, such as future climate projections, the LSTM might project different streamflow changes than the other models. A higher streamflow sensitivity to temperature changes by the LSTM compared to conceptual models has previously also been shown by Martel et al. (2025).”

Another concern relates to the importance of the different input features. Is temperature the most important predictor, or do other variables differ more between the cold and warm periods? The manuscript does not discuss precipitation changes, and I think a feature-importance/SHAP analysis is possible for the LSTM or hybrid model. It would be helpful to understand whether precipitation or PET, although changing less than temperature, may have a stronger influence on streamflow. Concerning the evaluation metrics are not very different from models to models during different period. just to confirm that different model performances are due to climate warming.

As shown in Figure 2b–f, temperature is the climate variable exhibiting the most pronounced change between the two periods, and this shift is closely mirrored by the decline in snowfall. Because snowfall is strongly correlated with temperature, the reduction in snowfall can be interpreted primarily as a temperature-driven effect. In contrast, precipitation and PET change only marginally between the cold and warm periods. Moreover, PET is generally low in the Alpine region, meaning that evaporation plays only a minor role in runoff generation.

Appendix C1 already includes a feature importance analysis. While this analysis mainly focused on static catchment attributes, we also incorporated the change in temperature as an additional feature. Our aim was to identify which catchment characteristics best explain differences in model performance across periods. The results indicated that specific discharge has the highest feature importance, but the temperature change also contributes to explaining some of the performance reductions.

A few minor comments

Line 1: Use consistent terminology: either “deep learning,” “deep-learning” (as an adjective), or “DL,” throughout the manuscript.

Thank you for noticing this detail.

We checked and now consistently use *“deep learning”*.

Line 1: Spell out “Long Short-Term Memory (LSTM)” on first use.

We have checked and LSTM is already spelled out both on first use in the abstract as well as on first use in the introduction.

The abstract is currently very conceptual. Please include key numerical results (e.g., NSE, KGE) to quantify performance. For example, Lines 10–12 mention that the LSTM performs best during the cold period but worse during the warm period, this should be supported with specific numbers.

From the abstract, the advantages of the hybrid model over the LSTM are not obvious. Lines 10–15 suggest that hybrid models have similar accuracy to LSTMs, please clarify the added benefit.

Thank you for this valuable comment. We will adapt the abstract accordingly.

The abstract has been revised to include clearer and more quantitative results, thereby explicitly demonstrating the added value of the hybrid model. The abstract now also incorporates key findings from the new generalizability analysis, further supporting the benefits of the hybrid approach.

> P1, L11-27: "[...] We find that the LSTM is the most accurate model overall (median NSE: 0.74-0.79), followed by the hybrid (median NSE: 0.68-0.71) and HBV model (median NSE: 0.53-0.56). However, the LSTM shows a stronger change in performance between the warm and cold period (NSE: -0.05 - -0.04) compared to the hybrid (NSE: -0.03 - 0) and HBV model (NSE: -0.03 - -0.02). When comparing observed with simulated streamflow signals, we find a higher sensitivity of the LSTM to changes in meteorological forcings (mainly temperature change) for mean flows (MQ: -8.5 - -12.6 %) and high flows (Q99: -7.0 to -13.4 %) than in observations (MQ: -4.6 - -9.4 %, Q99: -3.6 – -9.7 %). In contrast, the hybrid and HBV model show good agreement with observed change signals. These findings indicate that the LSTM, while being the most accurate overall, has a lower generalizability to warmer conditions than the other models. However, we also find that the models' capability to generalize can differ between catchment type. While all models generalize better in rainfall-dominated catchments, they all show larger but similar reductions in performance in snow-dominated catchments. This suggests that all models struggle to generalize to warmer conditions in snow-dominated catchments implying that better process-representation in the models might be needed to accurately capture the dynamics of snow-melt and -accumulation processes, which are highly sensitive to changes in temperature. We conclude that hybrid models effectively combine the strengths of LSTMs in having a high accuracy when predicting in ungauged basins, with the good generalizability under changes in climate from the conceptual hydrological models. This makes hybrid models a suitable choice for hydrological climate change impact assessments, particularly in ungauged basins."

Line 86: Please correct the citation formatting.

The citation format has been adjusted.

P2, L99: (Wi and Steinscheider, 2022, 2024) -> Wi and Steinschneider (2022, 2024)

The introduction is well written.

Thank you!

Line 153: If the Po River basin is not included in the analysis, it may be better not to mention it here (or clarify this later, as in Line 159).

We like this suggestion. We adapted the data description and now don't mention the "Po River" in line 153, but we keep the limitation of not including Italian stations in Line 159.

P6, L167-68: "specifically the **three** major river basins originating in the Central Alps: the Danube, Rhine, **and Rhône**."

Line 310: Please clarify the distinction between "in-sample HBV" and "regional HBV."

Thank you for highlighting the need for clarification. The "in-sample HBV" is trained locally for each catchment separately, which would be a classical use-case of this type of model, while the "regional HBV" infers model parameters from calibrated donor catchments. The latter setup

is better in line with the concept of LSTM training and allows us to evaluate the model's performance in "ungauged basins", which have not been used in training.

We have added multiple clarifications to Section 2.2.1.

- "In our analysis we distinguish between two different HBV setups: (a) HBV (in-sample) which was calibrated for each catchment separately, and b) HBV (regional) which infers model parameters from donor catchments and represents a model setup for ungauged basins. To obtain predictions for the HBV regional model, we employed the regionalization scheme outlined in Beck et al. 2016) with slight modifications."
- "If not explicitly mentioned otherwise, the standard HBV setup used for the model comparison is the regional HBV and is simply referred to as HBV."

Line 320: This relates to my major concern, how does precipitation change between periods and among different catchments?

The changes in precipitation and other meteorological variables are documented in Figure 2b–f. As shown there, the magnitude of precipitation change is minimal compared to the substantial shifts in temperature and snow fraction. Because snow fraction is strongly linked to temperature, its decline can be interpreted as a direct consequence of the warming signal rather than an independent change in hydrometeorological forcing. In other words, temperature is the dominant driver of climatic differences between the two periods, while other variables remain relatively stable. The stronger biases in mean flow by the LSTM, which we are referring to in line 320, are now also supported by the new analysis (see changes in streamflow metrics) showing stronger declining trends in mean flow by the LSTM compared to the hybrid and conceptual model and observations. This indicates a higher sensitivity to temperature changes of the LSTM. We will mention this link when we integrate the new figure.

We added a note on the climatic changes in Section 2.4.

> P10, L273-76: "These three scenarios (cold, warm and warmest) are distinct in terms of their climatic conditions and streamflow responses (Figure 2c-g). In terms of the median across catchments, the mean temperature increased by 1.4 °C between the cold and the warm period and by 2.0 °C between the cold and the catchment-specific warmest periods. The largest differences between the warm and warmest periods are observed in the temperature and snowfall variables, while potential evaporation and precipitation only show small changes."

Lines 315–319: The reported values are very close to each other, and they represent means or medians over hundreds of catchments. Could these differences fall within model uncertainty?

The differences are indeed not very large and non-significant, as reported in Figure 4. This could be due to different sources of uncertainty. Here we emphasise that there are not a lot of differences between the hybrid and the LSTM setups.

Section 3.4: In general, the hybrid and HBV models perform worse than the LSTM model. Is this due to limitations of HBV in snow-affected catchments, where LSTM may better learn snow–streamflow relationships? Does the hybrid model inherit these limitations from HBV, preventing it from outperforming the LSTM?

Partially yes. In the HBV setup there is no precipitation correction factor, which could minimize the influence of snow undercatch from the observations. In contrast, the LSTM model can learn that there is an input bias and could account for this. We discuss these limitations in the current version of the manuscript (l433-442).

Line 415: All models show higher performance for flood events than for drought or low flows. Is this due to the choice of objective function (NSE), which emphasizes high-flow periods?

Yes, this is a well-known challenge in hydrological modelling: high flows are generally "easier" to simulate than low flows. Importantly, the reduced model performance during low flow periods is not a consequence of using the NSE metric—similar patterns appear when applying alternative objective functions as shown in Bruno et al. (2024).

In general, low flow dynamics are more strongly influenced by factors such as human water management (e.g., hydropower operations, reservoir regulation, irrigation withdrawals, abstractions, or inflows from wastewater treatment) as well as groundwater contributions to streamflow (i.e., baseflow). These influences tend to play a larger relative role during dry periods than during high flow conditions. However, these processes are difficult to represent accurately, and detailed information about human interventions and groundwater dynamics is often unavailable. Consequently, both conceptual models and data-driven approaches typically struggle to reproduce low flow behavior with the same accuracy as high flow events.

We have included a short sentence on this in the discussion.

> P23, L488-91: "In contrast, all models have similar but lower accuracy for drought deficits, with the hybrid model being the most accurate of the regional models (Figure 5c). Differences in model accuracies for floods and droughts are common in hydrological modelling irrespective of the objective function used for training (Muñoz-Castro et al., 2026) and model performance often deteriorates under drought conditions (Bruno et al., 2024)."

Reviewer 2

Bohl et al. conducted the comprehensive benchmark of different types of models (purely data-driven, hybrid, and conceptual models) for the generalizability to warmer scenarios and extreme events. They found that hybrid models can be more robustly generalized to scenarios with distribution shifts compared to the LSTM and stand-alone HBV. The manuscript was well-written to follow clearly. I appreciate the comprehensive evaluation framework and the deep discussions provided in the paper and support this study to be published. I have some moderate/minor comments below mainly toward helping the manuscript further getting improved in the clarification of methodology and results.

Thank you very much for your positive feedback and your constructive and valuable comments. Please find our responses to your comments in blue following your comments in black.

I suggest that the authors have a table summarizing the details of all the evaluation experiments and scenarios, so that the readers can quickly review and refer to each experiment discussed, especially given that the study discussed different models for different types of generalization scenarios.

Thank you for this great suggestion. We added a schematic describing the methods choices (see a first version below). In addition, we adapted the text in Section 2.4 to make the information more accessible.

We have revised and expanded the description of the evaluation framework to improve readability and clarity. Further, we have added a schematic overview of the evaluation framework as a new subpanel in Figure 2. See below the new Figure 2 and its corresponding Figure caption.

See panel a) in the new Figure 2 and see Section 2.4 for an improved description of the evaluation framework.

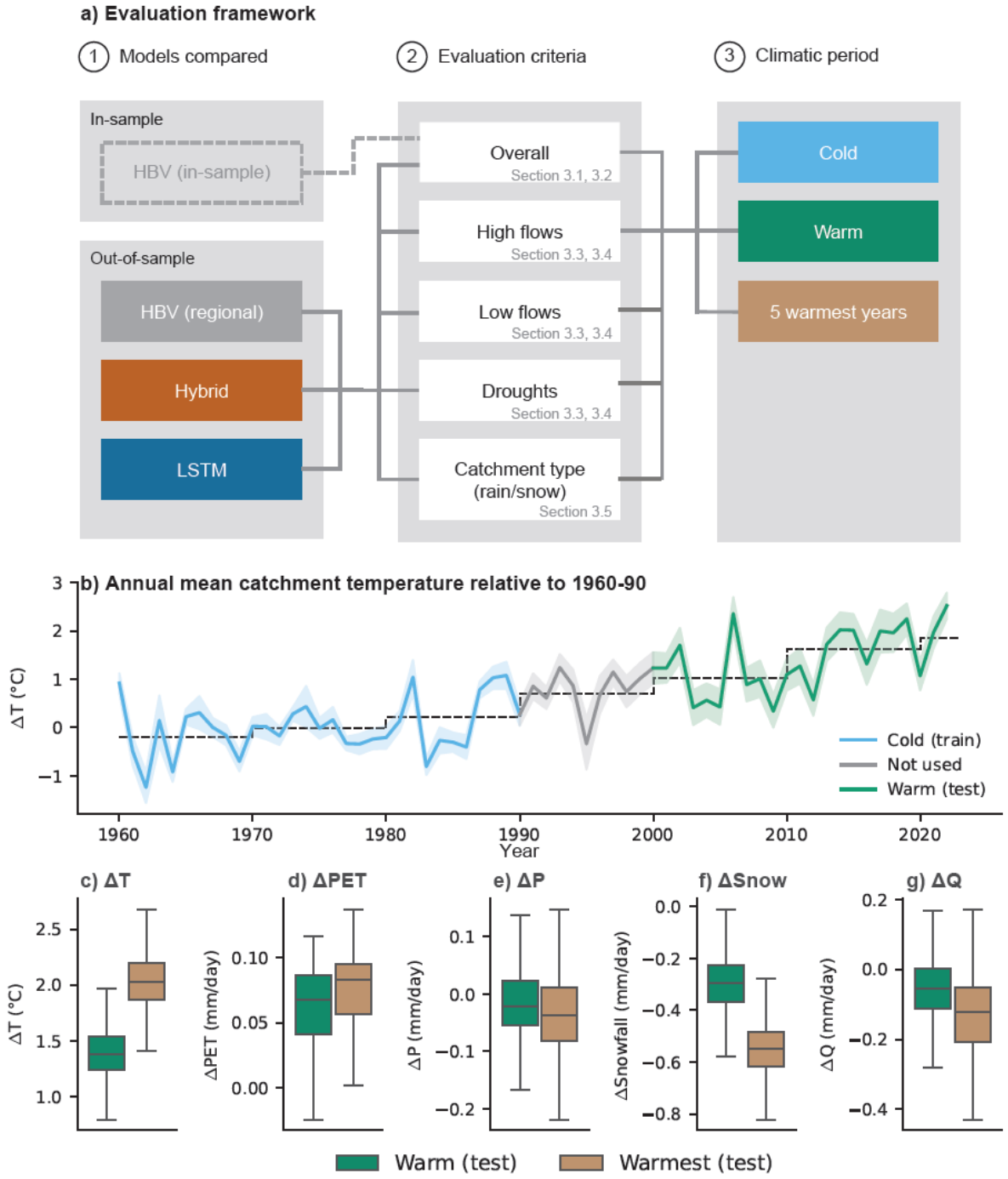

Figure 2: **Schematic of the evaluation framework and an overview of the climatic conditions in the three evaluation periods considered. (a) Overview of the different models, evaluation criteria and climatic periods used. (b)** Evolution of the yearly mean temperature relative to the 1960-1990 mean for each catchment, with the line indicating the mean over

catchments and the shaded area representing the standard deviation in each direction. The colors indicate the cold and warm period. The dashed lines indicate the decadal mean temperature relative to the 1960-1990 mean averaged over all catchments. **(c)-(g)** Climatic conditions during the warm and warmest evaluation period compared to the cold training period. The warmest period is defined as the 5 warmest years for each basin. For each variable, the catchment average of the cold period is subtracted from the catchment average of the evaluation periods. The boxes show the interquartile range, with the horizontal line indicating the median. The whiskers indicate the data points furthest away from the first and third quartile but still within 1.5 times the interquartile range from those quartiles. Outliers are not shown for improved readability."

Are the generalization tests for warm and warmest periods also for the predictions in ungauged basins? For example, the historical observations of those basins are not used during training? Please clarify.

Yes, exactly, we test the generalizability in both out-of-sample (ungauged) and out-of-time periods. We now clarify this in the new schematic (see above) as well as in the adapted Section 2.4.

We have revised and expanded the description of the evaluation framework in Section 2.4 to improve readability and clarity. We specifically added the following section to clarify that we perform our evaluation jointly out-of-time and out-of-sample.

> P10, L277-83: "In addition to the out-of-time evaluation, given by the three different periods, we also evaluate the models for the prediction of streamflow in ungauged basins (i.e. out-of-sample) in a total of 918 catchments. Specifically, we evaluated the models on the subset (fold) of catchments which was not used to calibrate or train the models. The combination of the out-of-time and out-of-sample evaluation is necessary to limit the influence of overfitting on the training data. This enables us to compare the model simulations across climatic periods without relying on training data. The HBV model was also evaluated in gauged catchments to evaluate the effectiveness of the regionalization scheme and provide a reference model typically used for climate change impact analyses. We refer to this model as the in-sample HBV as opposed to the regional HBV."

Does in-sample HBV mentioned throughout the manuscript represent the calibrated HBV in each individual catchment? Therefore, only the stand-alone conceptual HBV has in-sample prediction results, while all other results reported for the LSTM and hybrid models are for prediction in ungauged basins?

Yes, in-sample means that the HBV has been trained locally for each catchment, which means we can only test out-of-time performance, but not out-of-sample performance. The in-sample HBV is only used for the comparison of overall model performance in Figure 3. For all other comparisons between different models, the regional HBV, i.e. the HBV model making predictions for ungauged basins, are used. We adapted the text in Section 2.2.1 to clarify the distinction between HBV in-sample and HBV regional.

We have revised and expanded the description of the two HBV model setups in Section 2.2.1.

- "In our analysis we distinguish between two different HBV setups: (a) HBV (in-sample) which was calibrated for each catchment separately, and b) HBV (regional) which infers model parameters from donor catchments and represents a model setup for ungauged

basins. To obtain predictions for the HBV regional model, we employed the regionalization scheme outlined in Beck et al. 2016) with slight modifications."
- "If not explicitly mentioned otherwise, the standard HBV setup used for the model comparison is the regional HBV and is simply referred to as HBV."
- Please also see clarifications made in Section 2.4 (see responses in the comment above).

Line 175, my understanding is that the SWE observations used for the evaluation are also model-based data with uncertainties instead of ground-truth. Could the authors discuss how these data might impact the evaluation of the hybrid model simulated SWE?

Yes, your understanding is correct. The SWE data that we used as "observations" are also model based. Catchment-scale "ground-truth" data for SWE does not exist because SWE is only measured at specific point locations. To quantify catchment SWE, one has to rely on model output. We used two modelled SWE datasets that are considered to be reliable because they have been generated with national gridded meteorological datasets, which can be considered as the best estimate of local meteorology, and the SWE datasets partially used data assimilation to improve model results. Alternatively, we could have relied on SWE estimates from reanalysis products such as ERA5(-Land) or CERRA-Land. However, reanalyses often exhibit substantial SWE and snowfall biases in complex Alpine terrain (e.g., Monteiro and Morin 2023, https://doi.org/10.5194/tc-17-3617-2023), making them less suitable for our evaluation.

Differences between the Hybrid and HBV model and our "SWE reference" could stem from the difference in meteorology used to produce these SWE estimates; E-OBS for the hybrid and HBV model compared to two national and higher-resolution gridded products for Switzerland and Austria. We already discussed these limitations in the original draft in the discussion section. *"l.433-36 Our investigation of simulated SWE for the hybrid and HBV models revealed that both types of models systematically underestimate the water content of the snowpack (Figure 8). This indicates that there are biases in the forcing data and/or that the models are parametrising snow in a suboptimal way."*

I feel the used "LSTM-HBV" might not be an acronym accurate enough to represent the hybrid models developed in Feng et al. 2022 and 2023 which give a specific name called "differentiable hydrological models" or "δ (delta) models". LSTM-HBV reads like a loose coupling or post-processing type of models, which doesn't reflect the core of these hybrid models. Moreover, although the hybrid models use HBV as the base backbone, the frameworks also largely modified the original HBV structure.

We removed "LSTM-HBV" and now simply call it *"Hybrid model"* in the text and figure captions. In Section 2.2.3, we now clearly mention that hybrid refers to the delta models in Feng et al. 2022, 2023. In the introduction, we now have included the references (as suggested below) and include the term "*differentiable hydrologic models*".

We have harmonized and refined the description of the hybrid model throughout the manuscript.

- P1, Abstract: "[…] and (3) hybrid **(differentiable hydrologic models**) […] "
- P5, L149: "[…] a hybrid model (**differentiable hydrologic models, Feng et al. (2022a, 2023**))."
- P7, L205: "[…] a hybrid model, specifically the **differentiable hydrologic δ-HBV model**. "

- P9, L222: " We considered **the hybrid δ-HBV model** [...]"
- In figure/table captions (Figures: 3, 4, 5, 7, 8, 9, C1, C2; Tables: C2, C3, C4, C5) we changed the naming to "**hybrid** (**δ-HBV**)"
- In the text we now consistently use **"hybrid model"** instead of "LSTM-HBV".
- Section 2.2.3: Changed heading to **"Hybrid model".**

Please cite Feng et al., 2022 and 2023 in line 135 and 206 when referring to the "hybrid model" because the authors employed the hybrid models introduced in these previous studies.

We have included the following references in the introduction (i.e. line 135) *"(differentiable hydrologic models, Feng et al. (2022a, 2023)."*; as well as included Feng et al. 2023 in Section 2.2.3.
We have harmonized and refined the referencing of the hybrid model throughout the manuscript.

- P5, L149: "[...] a hybrid model (**differentiable hydrologic models, Feng et al. (2022a, 2023**))."
- P9, L226: "[...] proposed in the original **studies by Feng et al. (2022a, 2023)**, [...]"

NM7Q should be just one value for each year instead of time series based on the LFPB equations. Please modify the related use of "time series" when introducing the concepts.

Yes, this is correct, the NM7Q is a single value for each hydrological year. We assume that this comment refers to the use of "time series" on page 10. We rephrased this to:
P11: "[...] ***with simulated $\widehat{NM7Q}_y$ and observed $NM7Q_y$*** *for the hydrological year y."*

Figure 7, which catchment is this figure plotted on? or this is actually the mean across all catchments and days of the year? Please clarify.

These are seasonal average streamflow regimes in snow-dominated catchments. We will include *"average"* in the caption to clarify this.

Notice that Figure 7 has become Figure 8 in the revised manuscript. We changed the figure caption to, (P22, Figure8) "Seasonal **average** streamflow regimes [...]".

Line 409, I feel "it generalizes equally well..." can be a bit misleading given that the conceptual model's absolute performance is apparently lower than the other two models. Maybe a more accurate statement is the conceptual model has similar performance reduction to the hybrid model when generalized to warmer climate conditions?

We rephrased this to *"However, the conceptual model shows a similar small reduction in accuracy to the hybrid model indicating a similar generalizability to warmer climate conditions as the hybrid model."*

We have rephrased this.

> *P21, L473-75: "However,* ***the conceptual model shows a small reduction in accuracy similar to the one of the hybrid model, which indicates a similar generalizability to*** *warmer climate conditions as the* ***one of the*** *hybrid model"*

Line 413 the use of "the latter" is confusing here. I guess you refer to the conceptual HBV model but it's not clear.

The "latter" refers to both hybrid and HBV, but especially to the HBV.

We have clarified this.

> P22, L479: "[…] higher accuracy than the **HBV model** […]."

Line 417 is there a typo here? It seems the hybrid model is the most accurate in Figure 5c for drought volumes but you are saying HBV here.

Yes, it should say hybrid. We changed this.

We have clarified this.

> P23, L487-89: "**In contrast**, all models have similar **but lower accuracy for drought deficits**, with the **hybrid model** being the most accurate of the regional models (Figure 5c).

I am glad that the authors provide discussions on the potential reasons of underperformance in snow dominant catchments and the benefits of hybrid models over purely data-driven models for predicting untrained variables in line 435 and 456, respectively. Good jobs! These points are valuable and important to further think about.

Thank you very much!

Code availability: The authors of Feng et al., 2022 have released all their model codes in Zenodo publicly, while NeuralHydrology reimplemented these codes in their library. Therefore, credits should also be given to the original developers, such as in line 522 by noting, using the Python library that reimplemented the model codes in Feng et al., 2022, and citing the original Zenodo release.

Feng, D., Shen, C., Liu, J., Lawson, K., & Beck, H. (2022). differentiable parameter learning (dPL) + HBV hydrologic model. Zenodo. https://doi.org/10.5281/zenodo.7091334

You are absolutely right, credit should also go to the original developers. We now cite both Acuna Espinoza, as we used their implementation of the codes, and give appropriate credit to Feng, which is the basis for these codes. We adapted the text in the following way:

We extended the Code and data availability statement to explicitly acknowledge and credit the original developers of the hybrid model.

> *"[…] We specifically used the version provided by (Acuña Espinoza et al., 2025), since it includes an implementation of the hybrid model (differentiable parameter learning (dPL) + HBV hydrologic model) developed by Feng et al., 2022a. The code that we used is available at https://zenodo.org/records/14191623 (Acuna Espinoza, 2024) and the original code of the hybrid model at https://zenodo.org/records/7943626 (Feng et al., 2022b). […]"*

CC comment (Sascha Ruzzante)

This is a useful and timely paper. The testing of model generalizability follows best practices in the hydrologic literature, but I'd like to take this opportunity to ask exactly what is meant by 'generalizability'. Is it:

a) Which model has the highest accuracy in an unseen warm test period?

b) Which model has the smallest reduction (or largest increase) in accuracy when moving from

calibration (cold) to testing (warm) periods?

c) Which model most accurately simulates the change in hydrologic conditions between warm and cold periods?

Thank you for your comments. We define generalizability by conditions a + b. In our opinion, condition c alone is not a good proxy for generalizability, and can only be used as an extension of a and b.

We extended our generalizability framework by a definition similar to your proposed c).

P7, L203: "[...], and (4) their ability to simulated observed streamflow signals."

These three alternate definitions are subtly different and suggest different statistical tests. This study relies on definitions (a) and (b). In climate change projection studies, however, it is common to summarize results as a percentage change from historical conditions, for which definition (c) is the most relevant.

Ideally, all three conditions apply for a model "suitable" for climate change assessments. c) alone doesn't make a model generalizable. The model can get the change right for the wrong reasons. However, if conditions a+b are fulfilled, then it is likely that the change is also modelled correctly.

As an illustrative example: in a catchment, suppose the observed peak flow increases by 50%, from 100 cms to 150 cms, between the cold and warm periods. Models A and B give the following results:

Period	Observed	Model A	Model B
Cold	100	90	90
Warm	150	135	148
Change	50%	50%	64%

Model A has a persistent bias of -10%, and correctly predicts an increase in peak flows of 50%, while model B overpredicted the increase (64%). However, by definitions (a) and (b), we would select model B as the most generalizable since its accuracy in the warm period is highest and the accuracy improves from the calibration (cold) to the testing (warm) period.

We don't fully agree. According to definitions a+b we would still choose model A, because generalizable models should 1) have good initial performance, i.e. models A and B would be comparable (i.e. both have a -10% bias), and 2) the model performance should not change too much between periods (i.e. model A would meet this criteria with a constant bias of -10%, no change in performance). Following these criteria, a model that shows a significant increase in performance will also fail the stability criteria (b) and the change in performance would indicate that the model's performance is tied to external factors, e.g. temperature.

This example is relevant to Table C1, where (for example) the hybrid model is shown to have the best performance for DVPB in the warm period (4.9%), but this represents a large reduction from the cold period DVPB (9.5%). In comparison, the LSTM has the most stable DVPB across the three periods, as indicated at L358. In this case, it seems that the LSTM will predict the

change in DVPB best, and be most generalizable by definition (c). These numbers would, however, be more informative if compared on a catchment-by-catchment basis rather than comparing the median values.

In all figures, the catchment-by-catchment changes in performance are summarized in the boxplots. Summarizing the results in the form of boxplots was a pragmatic choice because we needed to find a way to quantify the overall behavior of the different models across the 918 catchments in our dataset. Alternatively, we could have shown scatterplots for specific pairs of models, which would have made it difficult to visualize differences between models.

I recommend including (maybe in an appendix) a comparison of the observed and simulated change in various hydrologic signatures between the cold and warm periods (eg., the mean annual flow, the mean monthly flow for each month, and the high flow, low flow, and drought metrics already calculated in the paper).

Thank you for these suggestions. We followed your recommendation and created a new figure quantifying the modelled changes in key streamflow metrics between the warm (warmest) and the cold period and comparing these changes to the observed ones. This allows us to assess not only whether model performance changes across periods (definitions a+b), but also whether each model appropriately captures the magnitude and direction of change in the respective metric in response to mainly changes in temperature (definition c). It seems that the conceptual and hybrid model can better capture the observed changes between the warm and cold periods. See figure below.

We added a new section (Section 3.4), the new Figure 6, adapted the abstract, discussion and conclusions to reflect the new analysis. Please see our response to Reviewer 1 for our detailed changes.

As a second point, the LSTM is found to generalize most poorly in the warmest catchments. To me, this makes sense, given that the LSTM is extrapolating most strongly in these catchments. In the colder catchments (warm period), the LSTM can learn from the warm catchments (cold period). For the warm catchments (warm period) there is no analogue set of catchments from which to learn. It might be worthwhile to mention this explanation alongside the explanations already given (L427-442).

We agree that it is important to comment on this point, which we already do in *"l.462-66: Further, model generalizability could be improved by including catchments from regions whose climate is warmer than that of the region of interest as additional training catchments for deep learning based models (Martel et al., 2025). This allows these models to learn from a more diverse set of climates, which could help improve generalizability. Studies testing this approach on LSTMs have found that such geographically extended models can provide more realistic projections under climate sensitivity analyses than regional LSTMs (Wi and Steinschneider, 2022, 2024)."*

Based on our new findings on trends we added the following to the discussion of the LSTM generalizability to warmer conditions.

We have extended the discussion section (4.1) on this:

> *P22, L479-84: "The lower generalizability of the LSTM compared to the hybrid and HBV model also holds when comparing observed and simulated streamflow signals. The LSTM simulations show a higher sensitivity to changes in meteorological forcings than*

observations (Figure 6), that is stronger declines in mean and high flows. This might indicate that when moving to even warmer climate conditions, such as future climate projections, the LSTM might project streamflow changes differently to the other models. A higher streamflow sensitivity to temperature changes by the LSTM compared to conceptual models has previously also been shown by Martel et al. (2025).”