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.

Below, we provide an example illustrating this idea by comparing relative changes in mean streamflow, mean high flow (annual 99th percentile), and mean NM7Q across different groups of catchments: all catchments, snow-dominated, and rainfall-dominated. Relative changes are given between the warm and warmest period compared to the cold period, respectively. We will add the figure and a description of the results in the revised manuscript. The results indicate that the LSTM seems to have a higher sensitivity to the imposed meteorological changes compared to the hybrid and conceptual model, i.e. this could be interpreted as a higher temperature sensitivity of the LSTM, which is in line with the results in Martel et al. 2025. In the revised version of the manuscript, we will also quantify the catchment-by-catchment agreement of these change signals.

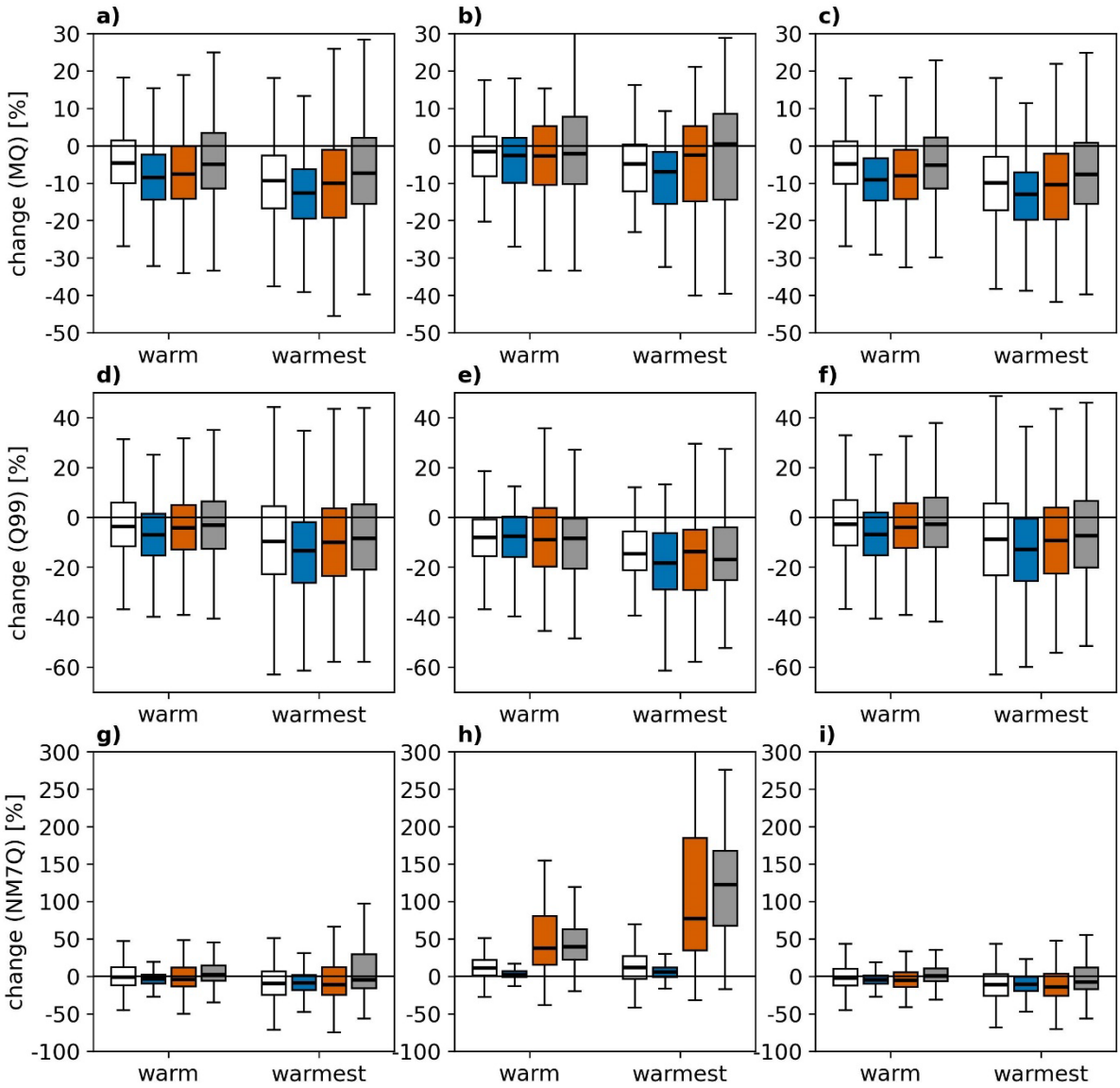

Figure x: Comparison of modelled and observed changes in streamflow metrics between the warm (warmest) and cold periods. Boxplots show catchment specific relative changes in (a-c) mean streamflow, (d-f) mean annual high flows (99th percentile), and (g-i) mean NM7Q between the warm (warmest) and cold period. The first column of panels (a, d, g) shows changes for all

catchments, the center column (b, e, h) for snow-dominated catchments (more than 30% snowfall fraction and minimal glacier contribution), and the right column (c, f, i) for rainfall-dominated catchments (less than 30% snowfall fraction). Colors indicate observations (white), LTSM (blue), Hybrid (orange), and HBV (grey).

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.

Line 86: Please correct the citation formatting.

The citation format has been adjusted.

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.

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 now clarify this when describing the HBV model 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, i.e. a model setup for ungauged basins. To obtain predictions for the latter (i.e. HBV regional), [...]"

"If not explicitly mentioned otherwise, the standard HBV setup which is used for the model comparison is the regional HBV and is simply labelled 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.

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.

"[…] whereas for drought volumes all models have similar but lower accuracy, with the hybrid model being the most accurate of the regional models (Figure 5c). Different accuracies for floods and droughts is a well-known problem in hydrological modelling irrespective of the objective functions (Muñoz-Castro et al., 2026) and model performance often deteriorates under drought conditions (Bruno et al., 2024)."

Bruno et al. (2024), https://doi.org/10.1016/j.jhydrol.2024.131023
Muñoz-Castro et al., 2026 : https://doi.org/10.5194/hess-30-825-2026