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.

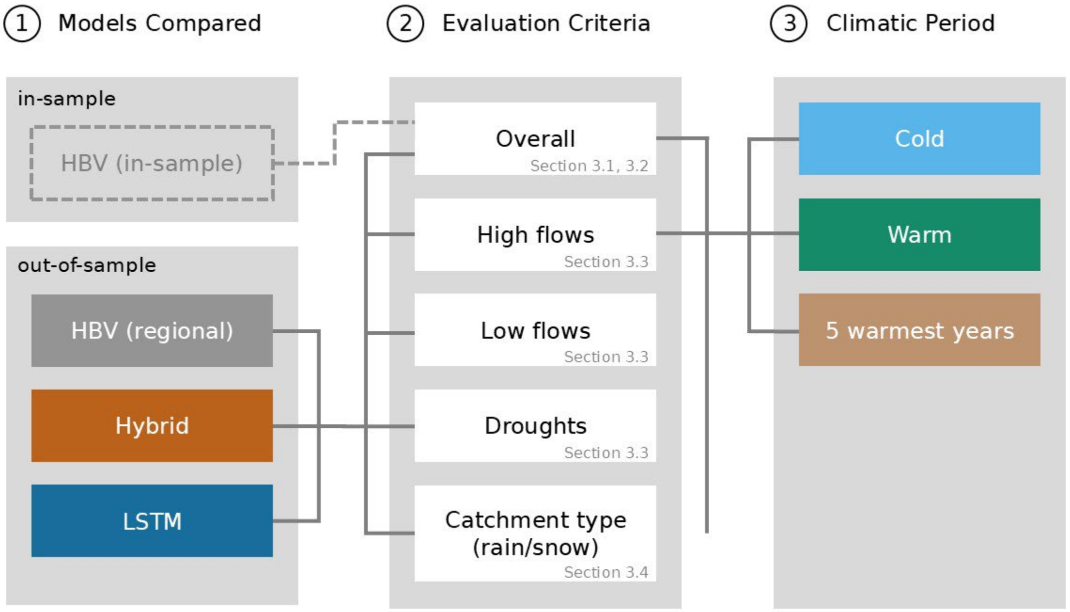

Figure X: Overview schematic of the different models, evaluation criteria and evaluation periods in this study. (1) Shows the four model setups used: HBV (in-sample), HBV (regional), Hybrid and LSTM. HBV (in-sample) is shaded as it is the only model setup that is calibrated per catchment individually and is only used for reference but is not part of the main evaluation. All other models are evaluated on unseen catchments in training (out-of-sample). The individual model setup and training procedures are detailed in Sections 2.2 and 2.3. (2) Indicates the evaluation criteria and the relevant sections. (3) Indicates the three climatic periods used for evaluation: cold (1960-1990), warm (varying periods in 2000-2023 with +1°C warming compared to the cold period), and warmest period (5 warmest years in 2000-2023). Models have been trained in the

cold period (1960-1990) and changes in performance are reported for the warm and warmest period compared to the cold period.

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.

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.

"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."

We also adapted the description in Section 2.4 to clarify this.

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*".

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.

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:
"[...] 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.

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."*

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 changed "latter" to *"hybrid and 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.

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 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)."