Response to Reviewer RC3

Comment n°35:

This paper investigates the question if LSTMs can be used for climate change impact prediction or rather: are they a valuable alternative to process-based hydrological models for climate change impact predictions in alpine environments? Or even more precise: are LSTM a valuable alternative to the selected process-based model (PREVAH)? Previous work has asked this question for other case studies and other models. From the paper as it is currently framed, it is unclear if this is simply a state-of-the-art case study or if it goes beyond the state-of-the-art. This having said, case studies can well desserve publication in HESS but given that the methods have all been used / have been developed in previous work, I think that the paper could make clearer what it contributes to the literature.

Co-authors' answer:

We thank Reviewer 3 for this comment. While LSTM-based hydrological models have been widely evaluated for historical runoff reconstruction, their stability and behaviour under climate change forcing remain comparatively little studied. Recent theoretical work has also suggested potential extrapolation issues in data-driven models (e.g. saturation effects (Baste el al., 2025)). In this study, we therefore provide a systematic assessment of the stability of LSTM-based runoff projections under the well established CH2018 climate scenarios in Switzerland and benchmark them against a well-established process-based modelling framework (PREVAH), which has been widely used for climate impact studies.

Comment n°36:

More importantly: as far as I see the reference simulations with the selected PREVAH model are problematic as reference simulations: these simulations are calibrated for some of the catchments; but for many other catchments, the model is somehow regionalized (we do not have information on this in the paper). I do not think that this is good practice: How can such a sample of two very different populations be used as a reference to evaluate a purely data-driven approach against? Furthermore, we see from Figure 3 that for many catchments, the reference model has an NSE value below 0.7 and even below 0.5: even if I cannot relate entirely to what 0.7 means for these catchments, 0.5 at daily time step is for sure a really bad performance for this climate, so these catchments should not serve as reference to compare the LSTM against. Furthermore, given the presented performance comparison, we simply know that the LSTM has a similar performance distribution. This is not informative: what if the LSTM does well for the the ones that PREVAH does not well and vice versa? If the focus of the paper is on seasonal signals only, we should at least expect a sensoring of the catchments / regions for which the seasonal signal is not well reproduced by a) by PREVAH, b) by LSTM, c) by both (which is very interesting!).

Co-authors' answer:

We acknowledge the point made by Reviewer 3 that calibrating hydrological models for individual stations can yield higher local precision. However, the objective of this study is to analyse spatially consistent runoff projections across Switzerland. Both modelling approaches considered here (PREVAH and the LSTM following Kraft et al.) are therefore used in a regionalization context suitable for large-scale assessments rather than locally calibrated setups. The modelling frameworks themselves have been developed and described in prior peer-reviewed studies (e.g. Viviroli et al., 2009, 2009a; Brunner et al., 2019 for PREVAH/Hydro-CH2018, and Kraft et al., 2024 for the LSTM framework), which provide details on the respective modelling setups, including the PREVAH parameter regionalization approach. In this context, the reported performance levels for predictions at ungauged locations are consistent with state-of-the-art regional hydrological modelling. This regionalization strategy is required to obtain spatially consistent parameter fields for ungauged locations and is standard practice in large-sample hydrological modelling studies.

Additional reference on the regionalization approach: Viviroli et al. (2009a).

Comment n°37:

Furthermore: we should not forget that we talk about streamflow simulations here but we do not see any actual simulations, not even for seasonal simulations: I think that it is of outmost importance to reconnect the data-based methods to actual data and hydrology. Otherwise we do not learn much about how well LSTMs perform for the selected hydroclimatic region.

Co-authors' answer:

We would like to clarify that the manuscript does present simulated runoff results, for example through projected long-term means and seasonal statistics. The focus on aggregated signals rather than day-to-day hydrographs is intentional, as climate-model–driven simulations represent plausible evolutions of climate forcing but are not expected to reproduce observed short-term discharge dynamics. The objective of the study is therefore to analyse projected runoff changes and their stability across models.

Comment n°38:

From a model development perspective, information is missing: how well does the PREVAH model, forced with climate data (rather than with observed precip & temperature) reproduce observed streamflow? This is a serious lack of information: the models are trained with observed data and then run with climate data; the divergence between the models for the future periode might be rooted at least partly in the divergence of their simulations for the observed period.

Co-authors' answer:

We thank the reviewer for this comment. As clarified in our response to a previous comment (RC3, comment 36), the PREVAH simulations used here originate from the Hydro-CH2018 dataset (Brunner et al., 2019b), where the modelling setup is described in detail. The underlying PREVAH model structure and calibration procedure are documented in earlier studies (e.g. Viviroli et al., 2009b for the calibration). The CH2018 climate forcing used for projections is bias-adjusted against MeteoSwiss gridded observations before being applied in the hydrological simulations. We will clarify this more explicitly in the manuscript.

Additional reference on the PREVAH calibration: Viviroli et al. (2009b).

Comment n°39:

Furthermore, the training input data set contains a gridded spatial rainfall product, which can be assumed to show very different quality for different catchments in an alpine environment. Do the climate scenarios have the same spatial resolution? And are the gridded climate scenarios debiased to the spatial rainfall product?

Co-authors' answer:

We thank the reviewer for this question. The CH2018 climate projections are bias-adjusted within the CH2018 framework using MeteoSwiss gridded observations (including RhiresD and TabsD) as reference and are provided on a 2 km grid. We will clarify this well-established processing chain more explicitly in the revised manuscript.

Comment n°40:

Otherwise, each of the models (PREVAH and LSTM) learns how to deal with the observation-based data set (i.e. how to translate it into observed streamflow), but then makes its own errors when applying what it has learned to a rather different data set. In this setting, what can you learn from such a comparison framework?

Co-authors' answer:

We thank the reviewer for this comment. Investigating how different modelling approaches behave when applied to out-of-sample climate forcing is precisely one of the objectives of this study. By comparing LSTM and PREVAH projections under the same CH2018 scenarios, we assess the stability of the data-driven approach relative to a process-based framework and highlight where their responses diverge (e.g., for extreme conditions under future climates).

Comment n°41:

Let's imagine: Model A is better than model B for the observed period if fed with observed data, but worse than B if fed with climate data. What do we learn from this? That model A was overfitted? Or that the climate data is not good? If now model A is close to model B for the future period, what does this tells us? That by chance, they both agree? I know that this is not the objective of the paper but given that it contributes to the question of how to use LSTM for climate change impact studies, these questions should be discussed.

Co-authors' answer:

We thank the reviewer for this comment. Interpreting agreement and divergence between modelling approaches is indeed an important aspect of climate impact studies. In general, consistent responses across models can increase confidence in projected signals, whereas differences highlight uncertainties in the modelling chains and call for cautious interpretation. We will discuss this aspect more explicitly and relate it to the broader practice of using multi-model ensembles to assess projection uncertainty.

Comment n°42:

Next: the hydrological process model certainly received PET as input for training. How was dealt with this for climate change simulations? And does the LSTM receive PET ? And if not, is this not unfair?

Co-authors' answer:

We thank the reviewer for this question. PREVAH is driven by a broader set of meteorological inputs used to compute evapotranspiration, whereas the LSTM is trained to reproduce observed discharge using precipitation and temperature together with static catchment attributes and glacier information as predictors. We acknowledge that the atmospheric forcing differs between the two modelling approaches, which may contribute to differences in projected runoff responses. This aspect will be briefly discussed in the revised manuscript.

Comment n°43:

Next: The LSTM is optimized with the mean squared error (MSE) between normalized simulated and observed discharge. This is certainly very different from the optimisation criterion used for the process model (MSE is rarely used). Why was not a similar / the same criterion used? And what can you learn from two models that are optimized based on different criteria? Besides: was the PREVAH model calibrated with some global optimisation method?

Co-authors' answer:

We thank the reviewer for this comment. The PREVAH calibration follows the procedure used in the Hydro-CH2018 framework (Brunner et al., 2019b), which builds on earlier PREVAH calibration studies (e.g. Viviroli et al., 2009). Model parameters were calibrated for 140 representative catchments using objective functions including volumetric deviation and benchmark efficiency (Viviroli et al., 2007; Schaefli and Gupta, 2007), and subsequently regionalized to the full model grid using kriging (Köplin et al., 2010). In this study, our objective is not to harmonize optimization criteria across models but to compare runoff projections produced by two established modelling frameworks. Differences in calibration strategies are therefore part of the structural differences between the approaches.

Comment n°44:

I believe that this paper needs to better frame the analysis framework, improve is A compared to B and what conclusions can be drawn in its methods section. Furthermore, we need more details on the model set up.

Co-authors' answer:

We thank the reviewer for this comment. The objective of this study is not to determine which modelling approach performs better, but to document and interpret differences between a data-driven LSTM and a process-based hydrological model when applied to climate projections. This comparison allows us to assess the potential suitability and stability of LSTM-based approaches for climate impact studies.

Comment n°45:

Abstract and elsewhere: what means "stable results" and how do you measure "physically credible"? "

Co-authors' answer:

We thank the reviewer for this comment. In the manuscript, "stable results" refers to the robustness and consistency of projected runoff responses across climate scenarios and modelling approaches. "Physically credible" refers to projected patterns that are consistent with established hydrological understanding for Switzerland (e.g. elevation-dependent responses, seasonal shifts related to snow and glacier processes). We will clarify these terms in the manuscript.

Comment n°46:

Intro: what do you mean by "These results highlight their robustness and scalability"? Does any of the references expliciteely assess the robustness of the models and if so, how is this defined in a hydrological modelling context? What is meant by scalability, what paper discusses this and how is it defined in a hydrological modelling context? Similar what is meant by "achieve strong generalization performance" in a hydrological context? The introduction has to be much more hydrology-specific.

Co-authors' answer:

We thank the reviewer for this comment. In the introduction, robustness refers to consistent model performance across different hydroclimatic conditions and catchments, scalability to the ability to apply a single model across many basins without local calibration, and generalization performance to predictive skill for unseen basins or time periods. These concepts are discussed in the cited literature, and we will clarify their meaning more explicitly in a hydrological context in the revised manuscript.

Comment n°47:

Intro: what do you mean by "Despite their success in present-day forecasting,"? Is this paper about present-day forecasting (forecasting = predicting current state based on previous state) or is it about continuous simulation (prediction) of hydrological states based on driver inputs (i.e. without feeding in the previous hydrological state)?

Co-authors' answer:

We thank the reviewer for this comment. In this sentence, “present-day forecasting” refers to the broader body of LSTM hydrology studies demonstrating strong performance under present-day conditions. The present study, however, considers continuous runoff simulation/prediction from meteorological driver inputs under climate forcing, not operational forecasting based on previous hydrological states.

Comment n°48:

The abstract states "Divergences are most pronounced in alpine and glacier-fed catchments, where runoff dynamics are more complex, yet the main governing patterns are captured. ": I would argue that the mean governing patterns in these environments can be capture by feeding only temperature into the LSTM; should this be toned down?

Co-authors' answer:

We thank the reviewer for this comment. The statement refers to the fact that, within the modelling framework used in this study, the LSTM reproduces the main regime-dependent runoff patterns observed across Switzerland, including in alpine and glacier-fed catchments. Assessing the relative contribution of individual predictors (e.g. temperature alone) was not an objective of this study, which focuses instead on comparing projected runoff responses between modelling frameworks. We note that hydrological projections in cryosphere-influenced catchments are generally associated with higher uncertainty due to their complex process dynamics (Addor et al., 2014).

Additional reference on the uncertainy of projections in cryosphere influenced catchments: Addor et al. (2014).

Comment n°49:

Section 2: "The dataset includes most of the variables used in Kraft et al. (2024) and combines those typically employed to force the spatially distributed model PREVAH" - can this be more precise: did PREVAH use similar data or might part of the divergence come from different reference data for parameter estimation?

Co-authors' answer:

We thank the reviewer for this comment. As noted above, we will clarify the respective model inputs and setups to better document differences between the LSTM and PREVAH frameworks. These differences in input data and forcing are part of the modelling-chain differences considered in this study.

Comment n°50:

Section 2: do the same dynamic glacier fractions feed into the LSTM and into the PREVAH simulations or do they receive different input? This is unclear in my view.

Co-authors' answer:

We thank the reviewer for this comment. Both modelling frameworks rely on the same underlying glacier evolution projections from Brunner et al. (2019b). In the LSTM setup, theses projections are aggregated to catchment-level glacier fractions used as a dynamic input, whereas in Hydro-CH2018/PREVAH they are integrated within the model grid and affect the glacier melt component. Thus, the same glacier projections are used but incorporated differently in the two modelling frameworks. We will add a brief clarification in the methodology to make this explicit.

Comment n°51:

Section 4: what is CH-RUN? This should be clearer somewhere

Co-authors' answer:

We thank the reviewer for this comment. CH-RUN refers to the LSTM-based runoff simulations presented in Kraft et al. (2024). We will add a short clarification at its first mention in the text.

Comment n°52:

What is "dynamical stability of projections", should this be defined in the methods?

Co-authors' answer:

We thank the reviewer for this comment. Here, "dynamical stability of projections" refers to the robustness of simulated runoff responses under climate forcing, i.e. whether projected patterns remain coherent with known hydrological dynamics when models are applied beyond the climate conditions represented in the training data.

Comment n°53:

The paper mentions that PREVAH is widely used; the model is however only unsed in Switzerland (given the references)

Co-authors' answer:

Here, "widely used" refers primarily to hydrological modelling and climate impact studies in Switzerland, where PREVAH is a well-established modelling framework. However, the model has also been applied in other regions, for example in Austria and the Peruvian Andes (e.g. Koboltschnig et al. (2008), Andres et al. (2014)). This national focus remains consistent with the scope of the present study.

Additional references on the application of PREVAH in other regions: Koboltschnig et al. (2008); Andres et al. (2014)

References

Addor, N., Rössler, O., Köplin, N., Huss, M., Weingartner, R., and Seibert, J.: Robust Changes and Sources of Uncertainty in the Projected Hydrological Regimes of Swiss Catchments, Water Resources Research, 50, 7541–7562, https://doi.org/10.1002/2014WR015549, 2014.

Andres, N., Vegas Galdos, F., Lavado Casimiro, W. S., and Zappa, M.: Water Resources and Climate Change Impact Modelling on a Daily Time Scale in the Peruvian Andes, Hydrological Sciences Journal, 59, 2043–2059, https://doi.org/10.1080/02626667.2013.862336, 2014.

Koboltschnig, G. R., Schöner, W., Zappa, M., Kroisleitner, C., and Holzmann, H.: Runoff Modelling of the Glacierized Alpine Upper Salzach Basin (Austria): Multi-Criteria Result Validation, Hydrological Processes, 22, 3950–3964, https://doi.org/10.1002/hyp.7112, 2008.

Viviroli, D., Mittelbach, H., Gurtz, J., and Weingartner, R.: Continuous Simulation for Flood

Estimation in Ungauged Mesoscale Catchments of Switzerland – Part II: Parameter Regionalisation and Flood Estimation Results, Journal of Hydrology, 377, 208–225, https://doi.org/10.1016/j.jhydrol.2009.08.022, 2009a.

Viviroli, D., Zappa, M., Schwanbeck, J., Gurtz, J., and Weingartner, R.: Continuous Simulation for Flood Estimation in Ungauged Mesoscale Catchments of Switzerland – Part I: Modelling Framework and Calibration Results, Journal of Hydrology, 377, 191–207, https://doi.org/10.1016/j.jhydrol.2009.08.023, 2009b.