Response to Reviewer RC1

Comment n°1:

Figure 7: Change legend: Median PREVAH projections are shown in solid blue line, but the legend shows a dashed blue line

Co-authors' answer:

Thank you, will be modified.

Comment n°2:

l.539 deeper deeper

Co-authors' answer:

Thank you, will be modified.

Response to Reviewer RC2

Comment n°3:

Section 2.5: The glaciers play a crucial role in the future runoff predictions, I think. Could you add some more detail here. E.g. which/how many of the basins have no glacier extent in summer by end of century and is this consistently applied in both LSTM and PREVAH?

Co-authors' answer:

We agree and will clarify Sect. 2.5. Glacier evolution is prescribed consistently for both modelling frameworks using the same catchment-level glacier-fraction time series. We will add a short summary of glacier retreat across catchments (e.g., number/share of basins with glacier fraction below a small threshold by end-century) and explicitly state that this forcing is identical for LSTM and PREVAH; thus, projected differences are not driven by different glacier-retreat assumptions but by differences in model structure.

Comment n°4:

Section 2.6 and 3.1: You compare these two models, and try to attribute differences - but I don't know where the models actually differ. I strongly suggest to give an overview (a table with a side-by-side comparison would work nicely) of the diverging input data and model structure of PREVAH vs the LSTM - observed meteo, cc data (and meteo parameters used), cal-val-test approach, glacier data and simulation approach in PREVAH, spatial representation, observational vs projection catchments, you train the LSTM on natural catchments - what about reservoirs in the projection catchments - are those included in PREVAH...?

Co-authors' answer:

We agree that the manuscript should more clearly document how PREVAH and the LSTM differ in terms of inputs, model structure, calibration/training strategy, spatial representation, and treatment of glaciers and human impacts. We will add a concise side-by-side overview of the key differences relevant for interpreting model divergences.

Comment n°5:

Section 4.3/4.4/4.7 I like the maps. However, I wonder how a scatter plot LSTM vs PREVAH Runoff (could add the corresponding r^2) would look like next to the maps (one dot=one catchment, x=LSTM runoff, y=PREVAH runoff, color of dot could be represented by observational, projection, selected catchments or the elevation band color or the ecoregion ... whatever you find most appropriate). This would give a better sense of the actual differences between the models and allow more intuitive diagnostics for why you see the differences.

Co-authors' answer:

We thank the reviewer for this suggestion. We agree that scatter plots can be useful diagnostics. In this study, however, we focus on spatially explicit maps and elevation-stratified analyses, as elevation is the dominant control on runoff regimes and model differences across Switzerland. These analyses already provide a clearer and more physically interpretable view of the LSTM–PREVAH differences than an aggregated catchment-wise scatter plot. Given the already high figure density, we retain the current visualisation strategy, but such a plot could be provided as supplementary material or in the response for completeness.

Comment n°6:

l.132ff I'd like to understand the bias adjustment and cc data better: Which "Swiss gridded observations" - are those the same that you forced the model with (chapter 2.2)? Did you conduct the QM yourself or does CH2018 provide that? I suggest to discuss the implications if the bias adjustment was done on a meteorology dataset different to the forcing dataset. Also, how did you prepare the cc data for the models: Spatial aggregation->Bias adjustment. Or Bias adjustment->spatial aggregation?

Co-authors' answer:

We thank the reviewer for this clarification request. The CH2018 climate projections used here are already bias-adjusted within the CH2018 framework (QM) using MeteoSwiss high-resolution (2 km) gridded observational analyses as reference, including TabsD and RhiresD for temperature and precipitation. These are the same observational datasets used to force the LSTM during training. No additional bias adjustment was applied in this study. Bias adjustment is performed at the grid level in CH2018, and the resulting fields are subsequently spatially aggregated to the catchment scale before being used as model input. We will clarify this workflow in the manuscript.

Comment n°7:

l.347ff, 413ff and l.529ff I think the key question here is: Why is this happening and which model is closer to reality? I know that you cannot answer this question as it's the future. This likely goes beyond your scope, but I wonder if it would help (you could add your thoughts to the discussion):

1. when comparing to observed data. I assume these future events are significantly outside your training data? If that wouldn't be too extreme, you could look at the test data of the LSTM (section 3.5) and extract the maximum/minimum precip periods and check the same events for PREVAH, and evaluate which model is closer to obs (this could help for your discussion in chapter 5.7)?

Co-authors' answer:

We agree that such an analysis could provide additional insight into model behaviour under extreme conditions. However, the suggested approach would require an event-based evaluation of model performance against observations, which is beyond the scope of this study. Our analysis

instead focuses on comparing how the two modelling frameworks respond to projected precipitation conditions under climate scenarios. As a lighter diagnostic, we may indicate whether projected atmospheric forcing partly exceeds the range observed during the training period, which could help contextualize potential extrapolation effects.

Comment n°8:

Section 2.1: Suggest to better explain in the text what the differently labelled catchments in figure 1 are used for. e.g. unclear what "second part of our analysis" means in the caption at this stage.

Co-authors' answer:

We agree and will clarify the text and Fig. 1 caption to better explain the role of the different catchment sets.

Comment n°9:

It is unclear why 96 minimally impacted catchments were chosen and then projected on 307 - you introduce a bias already (e.g. missing out on reservoir impacts to name one obvious point)?

Co-authors' answer:

We agree that this requires clarification. The LSTM is trained on minimally impacted catchments to avoid learning catchment-specific management effects (e.g. reservoir operations) and to isolate climate-driven runoff responses. The projections are then applied to the 307 Hydro-CH2018 catchments, which represent a semi-idealized set of medium-sized basins covering Switzerland and are widely used to provide spatially continuous runoff information in national-scale studies (e.g. Brunner et al., Kraft et al.). For regulated basins, the resulting projections should therefore be interpreted as "naturalized" runoff rather than managed discharge. We will clarify this in the manuscript.

Comment n°10:

In this sense, I don't see a value in section 3.1 without mentioning PREVAH. I think it is more important to understand the differences than explaining the LSTM in detail alone (E.g. I'd be interested in how PREVAH simulates glacier, snow, ET which is needed to understand the differences presented (e.g. in section 4.5)).

Co-authors' answer:

We agree. As noted above, we will add a concise side-by-side overview of PREVAH and the LSTM to better contextualize Sect. 3.1 and highlight the key differences relevant for interpreting the results.

Comment n°11:

l.220ff I like this cross validation methodology. I assume, per fold, you have 12 catchments that are validation, 12 test and 72 train. Is it correct that you took for valid: year 2016, 2017, 2018 of the 12 validation catchments; year 2019-2024 of the 12 testing catchments and the remaining years of the 72 training catchment for train? But what do the grey shades in Fig 2ab mean - do

they correspond? what is the light grey in b that is not part of a?

Co-authors' answer:

Yes, this interpretation of the cross-validation scheme is correct. The grey shades in Fig. 2a and 2b are currently not intended to correspond directly, which we acknowledge is confusing. We will revise the figure and caption to harmonize the colour coding across panels (train/validation/test) and clarify that white areas in panel (b) indicate missing observations.

Comment n°12:

l.242/369ff as far as I understand your LSTM model architecture is the same as Kraft et al. 2024 - with the only difference that you added dynamic glacier data. I'd have expected the LSTM would perform slightly better with this additional information. Do you have a (short) reasoning why it didn't (including Kraft et al's LSTM in the suggested table might help to explain this - see comment to 2.6)?

Co-authors' answer:

In the mentioned study (Kraft et al., 2024), the NSE is 0.76 versus 0.75 in our setup. We consider this to be within the uncertainty margin, i.e., the models demonstrate very similar performance. Differences in the evaluation setup may also contribute: while we evaluate temporal extrapolation by leaving out a continuous time range at the end of the available data, Kraft et al. evaluate reconstruction skill with splits within the training range and one period at the end. In addition, random splitting can have a strong impact on reported model performance. Finally, glacierized catchments are rare in the training dataset and may therefore have only a limited impact on overall median performance. We will clarify this in the manuscript and briefly discuss how the representation of glacierized catchments in the training data may affect the model's ability to learn glacier dynamics.

Comment n°13:

l.339 I am confused by this. Yes, fig 8 shows a stronger decline for PREVAH vs the LSTM, but I can't see that in fig 7. For 7b,c,d,f I even see the opposite.

Co-authors' answer:

The apparent discrepancy arises because Fig. 7 and Fig. 8 address different levels of aggregation. Fig. 7 shows individual catchment responses, where the relative LSTM–PREVAH signal can vary, whereas Fig. 8 summarizes aggregated elevation-band statistics across many catchments. The statement at l.339 refers specifically to the aggregated signal in Fig. 8. We will clarify this distinction in the text.

Comment n°14:

l.347ff, 413ff and l.529ff I think the key question here is: Why is this happening and which model is closer to reality? I know that you cannot answer this question as it's the future. This likely goes beyond your scope, but I wonder if it would help (you could add your thoughts to the discussion):

2. to attribute this to PREVAH's 'knowledge' of the driving processes (such as the additional climate data it gets to calculate ET) different constraints (or no constraints) for the glacier extent, the internal mass balance constraint that the LSTM doesn't have (could the application of a mass-conserving LSTM improve the situation)?

Co-authors' answer:

We agree that the differences between the models can be interpreted in light of their structural assumptions and constraints (PREVAH explicitly represents key hydrological processes, while the LSTM relies on data-driven relationships without explicit physical constraints). We will make this attribution more explicit in the discussion and link it to the model comparison overview (suggested table).

Comment n°15:

l.584 You earlier mention that this interpretation of low-flow performance requires caution and I don't see this adequately evaluated in your paper to merit mentioning this in the short conclusion.

Co-authors' answer:

We agree. Given the associated uncertainties, emphasizing low-flow performance in the short conclusion is too speculative. We will revise the conclusion and instead highlight complementarity in more robust aspects (such as multi-scenarios testing).

Comment n°16:

l.93 when does the data end/what was your cutoff? Mention the spatial resolution.

Co-authors' answer:

We agree and will clarify the temporal coverage (cutoff year, refer to the cross-validation scheme) and explicitly state the 2 km spatial resolution of the RhiresD and TabsD datasets in the manuscript.

Comment n°17:

l.96 PRISM and SYMAP - reference and spell out on first use

Co-authors' answer:

We will spell out PRISM and SYMAP and add references at first use.

Comment n°18:

l.102 how did you spatially average in detail? Extract entire cells or did you 'split' cells? If the catchment area is small and the grids large, you can introduce errors particularly in mountainous terrain.

Co-authors' answer:

The grid cells falling within each catchment polygon were averaged. This approach is consistent with the workflow adopted in previous studies using the same datasets. We will clarify this more explicitly in the manuscript.

Comment n°19:

l.116 Can you give a rational why you used topsoil information only?

Co-authors' answer:

Topsoil properties were used as they are most directly relevant for runoff-related processes and to limit model complexity and additional uncertainty from deeper soil layers. In addition, this choice follows the setup used in Kraft et al. (2024) / CH-RUN, which we replicate here to ensure methodological consistency. We will clarify this in the manuscript.

Comment n°20:

l.126 suggest to add that it is based on the CMIP5 framework

Co-authors' answer:

It will be added.

Comment n°21:

l.129 you earlier write that the resolution was 2km - how is the product downscaled from the 12-50km CORDEX resolution to 2km?

Co-authors' answer:

EURO-CORDEX simulations are dynamically downscaled to 12 or 50 km by the RCMs. In CH2018, the resulting fields are bias-adjusted and interpolated to a 2 km grid using MeteoSwiss gridded observations as reference (CH2018, 2018). We will clarify this in the manuscript.

Comment n°22:

l.148 Suggest to make it clearer whether you did any glacier simulations of if this was Brunner et al. 2019b work.

Co-authors' answer:

All glacier evolution simulations originate from the work of Zekollari et al. and are used in the Hydro-CH2018 dataset described by Brunner et al. (2019b). No glacier modelling was performed in this study; we only derive catchment-level glacier fractions from the provided gridded outputs. We will clarify this in the manuscript.

Comment n°23:

l.150 I think you mean section 3.2

Co-authors' answer:

Correct, thank you

Comment n°24:

l.185 in section 2.5 you cite Brunner et al. 2019b as the source for the glacier data

Co-authors' answer:

In this sentence we refer specifically to the glacier evolution simulations themselves, for which Zekollari et al. (2019) is the appropriate reference. Brunner et al. (2019b) is cited where the Hydro-CH2018 dataset is described as the data source.

Comment n°25:

l.201 I don't see the 24 tested alternatives in Kraft et al. 2025 section 3.6 - seems you employed the pre-print version of their approach?

Co-authors' answer:

Correct, this refers to the preprint version. We will update the section reference to Sect. 3.5.

Comment n°26:

l.251 fig 4: suggest to add the number of catchments per boxplot (n=x) in the caption.

Co-authors' answer:

It will be added, n=307 catchments.

Comment n°27:

l.261 I don't think DJF is a period where much melt dynamics is going on in Switzerland and JJA is not really low flow in most of the catchments (you mention this yourself in l.271-272)

Co-authors' answer:

We agree and will revise the wording: DJF will be described as influencing snow accumulation and subsequent melt, and JJA as capturing summer runoff variability, including low flows and residual melt contributions.

Comment n°28:

l.262 add that this is about the annual panel

Co-authors' answer:

It will be added.

Comment n°29:

l.307 I think the LSTM surpasses PREVAH by 2060, or?

Co-authors' answer:

You are correct. The LSTM surpasses PREVAH from around mid-century onward, and remains higher by 2100. We will revise the wording to reflect this more precisely.

Comment n°30:

l.327 fig 8 caption "(mm period^{-1} vs 1991–2020) for 2071-2100" I think you mean something like 2071-2100 - 1991–2020?.

Co-authors' answer:

We agree. The legend will be changed to "Runoff change (mm period^{-1} vs 1991–2020)" and the

caption to "for 2071–2100 relative to the 1991–2020 reference period under RCP8.5".

Comment n°31:

l.476-482 the language here suggests we are still in results. Suggest to rephrase.

Co-authors' answer:

We agree and will rephrase this passage to adopt a more discussion-oriented tone, while retaining the narrative logic linking regime dependence and elevation effects.

Comment n°32:

l.539 deeper deeper

Co-authors' answer:

Thank you, will be modified.

Comment n°33:

l.576 whether

Co-authors' answer:

Thank you, will be modified.

Comment n°34:

l.715, l.623 provide links to final paper?

Co-authors' answer:

Thank you, will be modified.

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 explicitely 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, these 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.