# 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.