

Author Comment (AC2) in response to Referee Comment 2 (RC2) on egusphere-2026-3798

Manuscript: "Generalization of Deep Learning Models to Ungauged Glacierized Basins: Evidence from Alpine, Patagonian, and North American Catchments" — M. Maharjan et al.

Referee: Anonymous (RC2). Final author comment, 2 October 2026.

We thank the referee for the positive assessment and for the concrete suggestions, nearly all of which we adopt. Where a point overlaps with RC1 or RC3 we note it, so that the three responses can be read together.

1. Shortening the manuscript

RC2: I think the paper would benefit from being shortened. Some paragraphs are too long and could be shortened to strengthen the main message. The figures are often overly complex, likely because the authors wanted to be comprehensive.

AR: We agree. Our replies to the individual suggestions follow.

RC2: Remove Experiment 4. I struggle to see how this experiment adds value compared to Experiment 3.

AR: Accepted. Experiment 4 will be moved to the Supplement with a short results paragraph, and the main text will state once why it is reported there. Referee 3 proposed omitting Experiment 3 instead, while noting that omitting Experiment 4 would also work. We retain Experiment 3 because it forms a matched pair with Experiment 2: both train on 227 basins and differ only in whether the A and M pools may draw basins from the more representative partitions, which is the comparison that separates representativeness from training-set size. The main text will therefore present Experiments 1, 2, 3 and 5. We retain the original experiment numbering in this response so that it matches the numbering used by all three referees.

RC2: Shorten the abstract to highlight the main hypotheses and results (e.g., articulate around the three aspects I mentioned above).

AR: Accepted. The abstract will be shortened and organised around the three findings the referee identifies: (1) without glacierized basins in the training pool skill is poor, and non-glacierized mountain basins recover only part of it; (2) the three architectures respond differently to the size and composition of the training pool; (3) an externally modelled glacier-runoff input adds little. Statements that rank the architectures will be worded in line with our response to RC1 on the unequal input sets (AC1, G1).

RC2: Add two hydrographs to the main manuscript (not all subpanels), and remove some of the tables, as well as maybe figure 7 (e.g., move it to the supplementary information). In my opinion, hydrographs are helpful for readers to understand the main differences between model setups.

AR: Accepted. The simulated hydrographs are currently in the Supplement only (Figs. S7–S10, one basin per glacier-coverage class). We will bring two into the main text, one low-glacier (<2 %) and one glacier-influenced (≥2 %) basin, matching the two-class stratification we adopt in response to RC1 and RC3, with a reduced set of panels. RC1 points out that a single basin cannot represent a class, so they will be labelled as examples rather than as representative basins. Fig. 7 belongs to the summer analysis and moves to the Supplement with it (comment 7).

RC2: I don't think figures 3 and 4 are both needed in the main manuscript.

AR: Agreed. Fig. 4 will be moved to the Supplement, as RC1 also requests.

RC2: *I would generally remove tables, especially when some of the figures already show some of the underlying numbers.*

AR: Accepted in large part. Table 5 leaves the main text together with Sect. 4.2 (comment 7). Tables 2 and 3 (also RC3) and Table 6 (also RC1 and RC3) will be moved to an appendix. Two tables remain in the main text: Table 1 (basin counts per glacier-coverage class) and Table 4, which is the numerical companion to Fig. 3 and will also carry Experiment 5.

RC2: *The discussion is too long. For example, the discussion from lines 693 to 727 could be drastically shortened. For example, you could mention only two aspects: the added value of comparing your results to a model with an explicit glacier module, and how the performance reported in this study aligns with existing studies.*

AR: Accepted. L693–L727 will be condensed to the two aspects the referee names. For the first, we will relate our Experiment 5 result to studies that represent glaciers explicitly within hybrid or DL frameworks (Aguayo et al., 2025; Kraft et al., 2025), which RC1 brought to our attention. For the second, we will keep the comparison with published large-sample skill (Feng et al., 2021; Feng and Gleason, 2024; Yang et al., 2025) in shortened form. The passage on forcings at L717–L727 will be rewritten in line with AC1, G1. The remaining Discussion paragraphs will be condensed as well (also RC3).

2. Reservoirs

RC2: *I think one aspect is missing in terms of data selection and the underlying reasoning: some catchments are likely greatly affected by reservoirs. Could seasonal and annual dam storage and release have a similar signal to glacier runoff? If the authors do not want to invest time into dividing the catchment sets into several other groups, I think it would be nice to at least discuss that point.*

AR: This is a fair point that the manuscript does not address. Caravan does not screen basins for human influence: basins were taken from the source datasets on the basis of available catchment boundaries, a licence permitting redistribution and a catchment area of 100–2,000 km² (Kratzert et al., 2023), and our own selection added only the requirement of a continuous streamflow record. Regulated basins can therefore be present in every partition, including G. Caravan distributes the HydroATLAS degree-of-regulation and reservoir-volume attributes for each basin (Linke et al., 2019), but these are not among the static attributes supplied to our models (Table S3), so regulation is neither screened for nor visible to the models.

Regulation can resemble a glacier signal, but not in one direction. Storage in the wet season and release in the dry season can mimic the delayed late-summer contribution of glacier melt, whereas in Alpine hydropower basins the storage of summer melt for winter release damps the melt-season signature. In both cases the observed hydrograph of a regulated glacierized basin partly reflects reservoir operation rather than glacier storage, and the models receive no information about it.

Within each trial the same test basins are used for every model and every training partition, so a regulated test basin enters all contrasts within that trial alike; what differs between partitions is the training pool, and we have not quantified how its regulated share changes from A to M to G. We will (i) state in Sect. 2.1.1 that no screening for regulation was applied, either in Caravan or by us; (ii) add a Discussion paragraph setting out the two effects above and identifying regulation as an unquantified source of error in the observed streamflow of both training and test basins; and (iii) recommend, as future work, screening or stratifying basins by the HydroATLAS degree-of-regulation attribute already distributed with Caravan.

3. Time evolution of glacier coverage

RC2: *I missed a discussion on the time evolution of glacier coverage. Could this impact the results? I did not see this discussed anywhere.*

AR: We will add this to the Discussion. Glacier extent enters the study in two ways. The G partition and the glacier-coverage classes are defined from RGI v6 outlines, a single snapshot that we hold fixed over 1988–2014. The glacier runoff used in Experiment 5 (Rounce et al., 2023) comes from PyGEM–OGGM, which evolves glacier geometry through time, so that input does reflect changing glacier cover. None of the models receives glacier fraction as a static attribute. We have not tested whether glacier change affects the results. It could matter in two places: a basin close to the 2 % boundary may change class over the period, and a trend in the melt contribution between the training and test periods is one form of the non-stationarity raised by RC3 (AC3, major comment 2). We will state both, and, following RC1, we will not extrapolate the results to future glacier states.

4. Citation placement

RC2: *In the introduction, many references are grouped at the end of long sentences. While this is commonly done, it is better to use citations that are in the correct place and related to the topic of the corresponding sentence. For example, the reference "Thébaud et al., (2026)" (line 160 and other instances) is not suited to the statement it refers to. Their study has nothing to do with "projections under future glacier change scenarios." I could not check all the references, but I suggest the authors verify that their references align with their placement.*

AR: The referee is right. We will therefore audit every citation confirming that it supports the specific statement it is attached to. The grouped citations in the Introduction will be broken up so that each follows the claim it supports.

5. Experiment 5 introduced too late

RC2: *The distinction brought by Experiment 5 needs to be made clearer much earlier. I struggled to understand how the glacier data were included in the study before the explanations of the experiments.*

AR: Agreed; RC1 makes the same request. Sect. 2.1.3 will state at the outset how the glacier data enter the study: monthly glacier runoff (Rounce et al., 2023) is aggregated per basin, disaggregated to daily values, and used only in Experiment 5, as an additional dynamic input for the LSTM and Graph WaveNet and added to precipitation for δ HBV. Experiment 5 will be introduced together with the other experiments at the start of Sect. 3, and its results will be shown in Fig. 3 and Table 4 alongside them.

6. Surface net solar radiation used for only two architectures

RC2: *Surface net solar radiation is used for only two of the three DL architectures (Table S2a). An explanation of why this is the case is needed.*

AR: The difference follows from the model structures and the configurations we adopted, and we will explain it in the Methods. δ HBV has no process that uses radiation: snowmelt is computed with a degree-day (temperature-index) routine and evapotranspiration from potential evaporation, so radiation reaches the model only indirectly, through the ERA5-Land potential evaporation. Radiation could in principle be supplied to the parameter-learning network, but the configuration we used (Feng et al., 2022) does not do so and we did not modify it. The LSTM and Graph WaveNet were given the ERA5-Land variables of their source configurations (Kratzert et al., 2019; Sun et al., 2021), and we retained surface net solar radiation because shortwave radiation is a primary driver of snow and ice melt in glacierized basins (L53).

The radiation difference is part of the broader input difference that RC1 identifies as the main issue (AC1, G1): the two input sets are not nested, and the statement at L180 that all models received the same forcing is incorrect and will be removed. In response, we retrained the LSTM and Graph WaveNet for Experiment 1 on the δ HBV input set, i.e. without net solar radiation, snow water equivalent or surface pressure, for all three training partitions with 10 trials each (AC1, G1, Fig. R1). The NSE distributions of the LSTM were nearly unchanged, and those of Graph WaveNet changed only slightly and not consistently in sign; the differences between architectures and the effect of the training partition were the same as with the original inputs. Radiation, and the other inputs that δ HBV does not receive, therefore do not explain the differences between architectures in Experiment 1. We will report the re-run in the Supplement and state the input set of each model in Sect. 2.1.1.

7. Summer-month evaluation not introduced in Sect. 3.6

RC2: *Model efficiency is assessed for summer months in Section 4.2. However, I could not find any reference to this in Section 3.6. The reasoning behind this assessment should be made explicit.*

AR: The omission is ours: the reasoning, that glacier melt contributes most to streamflow in the melt season, appears only in Sect. 4. RC1 identified two further problems with this section: the June–August window is wrong for the Southern Hemisphere basins, and sub-period NSE is biased low by construction (Klotz et al., 2024). We will therefore move Sect. 4.2 (Table 5, Figs. 5 and 7) to the Supplement in corrected form, with a hemisphere-aware melt-season window and a normalised error metric in place of NSE. The window and its motivation will be defined in Sect. 3.6, and the Results will point to the Supplement in one sentence.

8. Signature-based metrics; NSE versus KGE in glacierized regimes

RC2: *I think this study would have benefited from more "signature-based" metrics in glacierized catchments. In such catchments, I would expect good KGE values to be easily reached because streamflow is highly autocorrelated. However, I would expect low NSE values because it is difficult to beat the benchmark; the mean is already a good predictor of streamflow. It would be interesting to see some discussion on this topic and how it could influence the results.*

AR: We agree that the benchmark implied by each metric matters for interpreting our results, and we will add a discussion of it. We read the mechanism differently, however, and this changes what follows for our results. NSE compares a simulation with the mean of the observations over the evaluation period. In strongly seasonal glacial and nival regimes the observations vary widely around that mean, so the mean is a weak benchmark, and NSE tends to be high for any simulation that reproduces the seasonal cycle (Schaepli and Gupta, 2007); RC3 raises the same point. The benchmark that is hard to beat in these basins is the mean seasonal cycle, which NSE does not use. KGE is raised by the same seasonality through its correlation term, and its values are not on the same scale as NSE: a KGE of about -0.41 corresponds to the mean-flow benchmark (Knoben et al., 2019).

We will (i) define NSE and KGE formally in Sect. 3.6; (ii) add a Discussion paragraph making these points, citing Schaepli and Gupta (2007) and Knoben et al. (2019), and cautioning against reading absolute NSE values as skill beyond the seasonal cycle; and (iii) state for each experiment whether KGE agrees with NSE, using the KGE results already in the Supplement, which move to an appendix at RC1's request. We will not add signature-based metrics in this revision: all three referees ask for a shorter paper, and the question the paper addresses, the effect of training-pool composition on skill, is answered by the distributions of NSE and KGE over the repeated trials. A signature-based evaluation will be named as future work.

Summary of changes in response to RC2

- Scope: Experiment 4 to the Supplement; Experiment 5 introduced in Sect. 2.1.3 and Sect. 3 and shown in Fig. 3 and Table 4.
- Length: abstract rewritten around the three findings; L693–L727 condensed to two points; remaining Discussion paragraphs condensed.
- Figures and tables: two example hydrographs brought into the main text; Fig. 4 and the corrected Sect. 4.2 (Table 5, Figs. 5 and 7) to the Supplement; Tables 2, 3 and 6 to an appendix.
- New Discussion material: reservoir regulation, with Sect. 2.1.1 stating that basins were not screened for it; time evolution of glacier cover; benchmark dependence of NSE and KGE.
- Metrics: NSE and KGE defined in Sect. 3.6; melt-season window defined in Sect. 3.6.
- Inputs: L180 statement removed; input sets listed; common-input re-run of Experiment 1 reported in the Supplement (AC1, G1).
- References: audit of existence and placement; Thébault et al. (2026) removed from statements on future glacier change.

References

- Aguayo, R., Zekollari, H., Hanus, S., Baez-Villanueva, O. M., Mendoza, P. A., and Maussion, F.: Hybrid glacio-hydrological modeling reveals contrasting runoff changes in western Patagonia over the 21st century, *Earth's Future*, 13, e2025EF006442, <https://doi.org/10.1029/2025EF006442>, 2025.
- Feng, D. and Gleason, C. J.: More flow upstream and less flow downstream: the changing form and function of global rivers, *Science*, 386, 1305–1311, <https://doi.org/10.1126/science.adl5728>, 2024.
- Feng, D., Beck, H., Lawson, K., and Shen, C.: The suitability of differentiable, learnable hydrologic models for ungauged regions and climate change impact assessment, *Hydrol. Earth Syst. Sci. Discuss.* [preprint], <https://doi.org/10.5194/hess-2022-245>, 2022.
- Feng, D., Lawson, K., and Shen, C.: Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data, *Geophysical Research Letters*, 48, e2021GL092999, <https://doi.org/10.1029/2021GL092999>, 2021.
- Klotz, D., Gauch, M., Kratzert, F., Nearing, G., and Zscheischler, J.: Technical Note: The divide and measure nonconformity – how metrics can mislead when we evaluate on different data partitions, *Hydrology and Earth System Sciences*, 28, 3665–3673, <https://doi.org/10.5194/hess-28-3665-2024>, 2024.
- Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, *Hydrology and Earth System Sciences*, 23, 4323–4331, <https://doi.org/10.5194/hess-23-4323-2019>, 2019.
- Kraft, B., Schirmer, M., Aeberhard, W. H., Zappa, M., Seneviratne, S. I., and Gudmundsson, L.: CH-RUN: a deep-learning-based spatially contiguous runoff reconstruction for Switzerland, *Hydrology and Earth System Sciences*, 29, 1061–1082, <https://doi.org/10.5194/hess-29-1061-2025>, 2025.
- Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward improved predictions in ungauged basins: exploiting the power of machine learning, *Water Resources Research*, 55, 11344–11354, <https://doi.org/10.1029/2019WR026065>, 2019.
- Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan – a global community dataset for large-sample hydrology, *Scientific Data*, 10, 61, <https://doi.org/10.1038/s41597-023-01975-w>, 2023.
- Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., Tan, F., and Thieme, M.: Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution, *Scientific Data*, 6, 283, <https://doi.org/10.1038/s41597-019-0300-6>, 2019.

- Rounce, D. R., Hock, R., Maussion, F., Hugonnet, R., Kochtitzky, W., Huss, M., Berthier, E., Brinkerhoff, D., Compagno, L., Copland, L., Farinotti, D., Menounos, B., and McNabb, R. W.: Global glacier change in the 21st century: every increase in temperature matters, *Science*, 379, 78–83, <https://doi.org/10.1126/science.abo1324>, 2023.
- Schaefli, B. and Gupta, H. V.: Do Nash values have value?, *Hydrological Processes*, 21, 2075–2080, <https://doi.org/10.1002/hyp.6825>, 2007.
- Sun, A. Y., Jiang, P., Mudunuru, M. K., and Chen, X.: Explore spatio-temporal learning of large sample hydrology using graph neural networks, *Water Resources Research*, 57, e2021WR030394, <https://doi.org/10.1029/2021WR030394>, 2021.
- Thébault, C., Knoben, W. J. M., Addor, N., Newman, A. J., and Clark, M. P.: Varying the combination of hydrological models in time and space: toward a more accurate representation of streamflow in large-sample hydrology, *Water Resources Research*, 62, e2025WR042272, <https://doi.org/10.1029/2025WR042272>, 2026.
- Yang, Y., Feng, D., Beck, H. E., Hu, W., Abbas, A., Sengupta, A., Delle Monache, L., Hartman, R., Lin, P., Shen, C., and Pan, M.: Global daily discharge estimation based on grid long short-term memory (LSTM) model and river routing, *Water Resources Research*, 61, e2024WR039764, <https://doi.org/10.1029/2024WR039764>, 2025.