

Author Comment (AC3) in response to Referee Comment 3 (RC3) on egusphere-2026-3798

Manuscript: "Generalization of Deep Learning Models to Ungauged Glacierized Basins..." — M. Maharjan et al. Referee: Anonymous (RC3). Final author comment, 2 October 2026.

We thank the referee for the careful reading and for recognising that the manuscript answers its questions comprehensively. We agree that it is too long and we accept most of the shortening suggestions. Where the referee's suggestions differ from those of RC1 or RC2 we explain the choice we have made.

Major comments

Major comment 1 — Length

RC3: *The manuscript is currently too long for the message it aims to convey.*

AR: Agreed. Our replies to each suggestion follow.

RC3: *Omit experiment 3: with experiment 1 as the baseline, experiment 2 and 4 deviate from it by respectively reducing the training dataset size and by excluding glacierized basins. I do not get the impression that experiment 3 adds information beyond 2 and 4. Referee #2 recommends omitting experiment 4, which I believe would also work. The fact that experiment 4 is not referenced in the discussion speaks for this option.*

AR: We will take the second option and move Experiment 4 to the Supplement, with a short results paragraph and a one-sentence pointer in the main text. 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. Experiment 4 differs from Experiment 1 in the size of the pools as well as their composition (A falls from 2,789 to 1,825 basins, M from 964 to 737) and rests on 10 trials rather than 100, so it is the weaker of the two tests of representativeness. The main text will present Experiments 1, 2, 3 and 5.

RC3: *Omit the annual NSE and keep only the summer NSE: since the aim is to improve glacier-influenced streamflow, evaluating only the summer months should be sufficient. I do not get the impression that the results are very different among the two metrics.*

AR: We would like to keep the full-period NSE and move the summer analysis instead, for three reasons. First, RC1 showed that our June–August window is wrong for the Southern Hemisphere basins and that NSE computed over a sub-period is biased low by construction (Klotz et al., 2024), so the summer analysis cannot carry the paper as it stands. Second, the paper asks whether the models generalize to ungauged glacierized basins, which concerns the whole regime, including the snowmelt rise and low flows, and the published studies we compare against report full-period skill. Third, as the referee observes, the two sets of results lead to the same conclusions, so the shortening is achieved either way. Sect. 4.2 (Table 5, Figs. 5 and 7) will move to the Supplement in corrected form, with a hemisphere-aware melt-season window and a normalised error metric in place of NSE.

RC3: *Figure 4 could be reduced to just the G training dataset.*

AR: Fig. 4 will be moved to the Supplement in full, as RC1 and RC2 both consider it redundant with Fig. 3.

RC3: *The High and Very High glacierized basin categories could be merged, considering that they contain only 4 and 2 basins respectively.*

AR: Agreed, and we will go one step further, since a merged class of six basins would still be too small to support comparison. As RC1 notes, the analysis in Sect. 4.3 in practice contrasts low-glacier basins with the rest. We will adopt two classes throughout: low (<2 %, 188 basins) and glacier-influenced (≥ 2 %, 95 basins), with counts stated in Table 1 and in every figure caption. The finer bins will be kept in Table 1 as a description of the sample only.

RC3: *The introduction and discussion chapters are well-structured and comprehensive, but I believe each of the paragraphs can be condensed. I would also omit the part on the computation time in the discussion.*

AR: Both accepted. Every Introduction and Discussion paragraph will be condensed; the computation-time material and Table 6 will be moved to an Appendix.

RC3: *Figure 2 is informative but could be reduced in size.*

AR: Accepted; Fig. 2 will be redrawn more compactly; the panel for Experiment 4 moves to the Supplement with that experiment.

RC3: *Tables 2 and 3 can be moved to the appendix.*

AR: Accepted. Because Table 2 lists the model inputs, the input sets will also be stated in the Methods text (see the comment on ERA5-Land SWE below).

Major comment 2 — Temporal splits and stationarity

RC3: *The k-fold sampling currently considers only basin splits, while the temporal training/testing split is kept constant across all experiments with 1988-2005 as the training period and 2008-2014 as the test period. I would recommend reducing the number of basin splits (currently 100) and adding several temporal splits where training, validation and testing periods are shuffled. Considering climate change and the occurrence of decadal droughts in two of the study regions (Garreaud et al., 2025; Hogan and Lundquist, 2024), there is a considerable risk that the assumption of stationarity between training and testing periods is violated. Testing for different combinations of training and testing periods would make the results more robust in this sense. As a side note, it is unclear from the manuscript why the years between 2005 and 2008 are not used, or why 1988 and 2014 define the lower and upper boundaries.*

AR: We agree that a single temporal split leaves the results exposed to non-stationarity, and we thank the referee for the two references, which we will cite. However, our experimental design holds the temporal split fixed deliberately: the question of the paper is how spatial/hydroclimatic composition of the training pool affects generalisation, and varying the period at the same time would confound that signal. A full temporal re-sampling would require retraining all models for every experiment and is not feasible within the final-response period. We will therefore state the fixed split and its consequences explicitly in Sect. 3 and in the limitations. The contrasts we report are between training pools evaluated over the same test period, so a shift between the two periods is common to every configuration we compare; we have not shown that it affects them equally, and we will say so. Temporal re-sampling, with fewer basin splits as the referee suggests, will be named as the first item of future work.

AR: On the side note: the reason we chose 1988 as the lower temporal boundary was because it is the earliest year with continuous streamflow records across the retained basins. We left out 2005-2008 as a validation period for early stopping as a buffer between training and test. We will state this in Sect. 3.

Minor comments

RC3: *I would recommend using a minimum glacier cover before considering a basin glacierized. It is unclear from the manuscript what the exact distribution of the glacier percentages among the 283 glacierized basins is, but I suspect that a number of them have such low percentages that the glacier runoff ratio is negligible.*

AR: The referee's suspicion is correct: 188 of the 283 G basins (66 %) have less than 2 % glacier cover (Table 1). We will add a histogram of glacier fraction to the Supplement and report its quantiles in Sect. 2.1.1. Redefining G with a minimum cover would change the training pools and the test sets of every experiment, and we will not do that in this revision. Instead we will state that G comprises the high-mountain basins that intersect any RGI glacier outline, and we will report the results separately for the low (<2 %) and glacier-influenced (≥ 2 %, 95 basins) classes, so that readers can see whether the findings hold for the basins the referee would regard as glacierized. A minimum-cover definition will be recommended as a refinement.

RC3: *The authors might consider relaxing the requirement of full streamflow data availability over the study period. At least for the LSTM, Gauch et al. (2025) have shown different effective strategies of dealing with missing data. This could increase the number of training basins, especially in the currently underrepresented Chilean Andes. However, this might not be so straightforward for δ HBV and GNN.*

AR: We agree this would increase the Chilean sample in particular (RC1 made the same observation), and we will cite Gauch et al. (2025) in the Discussion as the way to do it. Because the strategies differ by architecture and would change the basin sample for all experiments, we will keep the present selection and note this as future work, together with the use of Caravan v1.3, which adds further glacierized basins.

RC3: *The disaggregation of glacier runoff from monthly to daily timescale (L226-249) is done in a very similar way to the work of Wiersma et al. (2022). However, they use a Tbase of -5 °C and use a calibrated scaling factor ... ERA5-Land forcing is known to be biased in mountain areas and therefore possibly calls for a different threshold. Please justify your choices and mention any validation against observed glacier runoff data.*

AR: Thank you for the reference; we will cite Wiersma et al. (2022) and compare our scheme to theirs. We used Tbase = 0 °C as the conventional physical threshold and did not calibrate a scaling factor or a threshold, because the disaggregation only redistributes a fixed monthly total within each month, and no observed daily glacier-runoff data are available for our basins to calibrate or validate against. We will state this, note that the catchment-mean ERA5-Land temperature is likely warm-biased relative to the glacier (RC1 raised the same point), and add that the coarseness of the disaggregation is a plausible reason for the small effect of the glacier-runoff input in Experiment 5.

RC3: *A definition of the exact summer months used for the NSE calculation is missing.*

AR: The window is June–August, which is given only in the heading of Sect. 4.2 and, as RC1 points out, is wrong for the Southern Hemisphere basins. The corrected, hemisphere-aware window will be defined in Sect. 3.6 (see Major comment 1).

RC3: *Why is the HydroAtlas glacier fraction not included as one of the Caravan static attributes (gla_pc_sse)?*

AR: gla_pc_sse is not among the static attributes supplied to the models (Table S3). Glacier fraction in this study is computed from the RGI v6 outlines and is used only to define the G partition and to stratify the results; it is not a model input. We will state this in Sect. 2.1 and note that adding glacier fraction as a static attribute is an obvious test for future work.

RC3: *Why is ERA5-Land SWE included as a model input? ... it can still show strong biases in mountain areas (Lundquist et al., 2026). I suspect that the data-driven models are capable of accounting for the seasonality of snow and glacier processes from the meteorological inputs*

without relying on the modeled ERA5-Land SWE output. Its use as an input should be justified in the manuscript if it is kept.

AR: ERA5-Land SWE is part of the standard Caravan forcing set used in the LSTM and Graph WaveNet source configurations, and we originally kept it for that reason rather than on the basis of a test. We agree that it is a modelled product with known mountain biases, and we will cite Mudryk et al. (2024) and Lundquist et al. (2026) to say so. We have since tested the referee's expectation directly. In response to RC1's main comment on unequal input sets (AC1, G1), we retrained the LSTM and Graph WaveNet for Experiment 1 on the δ HBV input set (precipitation, temperature and potential evapotranspiration), i.e. without SWE, net solar radiation or surface pressure, for all three training partitions with 10 trials each. The NSE distributions of the LSTM were nearly unchanged, and those of Graph WaveNet changed only slightly and not consistently in sign (AC1, Fig. R1). This supports the referee's expectation that the DL models can represent the seasonality of snow and glacier processes from the meteorological inputs alone, at least in this setting. We will report the re-run in the Supplement and use it in Sect. 2.1.1 to justify retaining SWE in the main experiments: it is part of the standard deployed configuration and, on this evidence, does not drive the results. Experiments 2–5 were not re-run without SWE, and we will state this.

RC3: Fig. 1 misses Danish catchments. Either add the catchments to the figure or leave CAMELS-DK out of the study.

AR: The Danish basins will be added to Fig. 1. We will also make clear (as RC1 requested) that CAMELS-DK contributes only to the A partition, since it contains no glacierized or high-mountain basins.

RC3: L198: define the minimum mountain cover for a catchment to be considered a mixed high mountain catchment.

AR: M comprises the A basins whose polygon intersects the high and scattered-high mountain classes of Karagulle et al. (2017), identified by a spatial join; no minimum areal fraction was applied. We will state this at L198, together with its consequence that a basin with a small mountain share counts as M.

RC3: Fig. 6: mention in the caption what the colors represent.

AR: Accepted.

RC3: Figs 3-5: either add the asterisk to the figure or write the final caption sentence in default formatting.

AR: Accepted; the asterisk will be added to the figures.

RC3: L402: refer to the supplement for the KGE and RMSE results. I would argue against showing RMSE results, as NSE and RMSE are both squared residual metrics.

AR: We will refer to the KGE results (to be moved to an Appendix, as RC1 requested) at L402 and we will drop the RMSE table from the main comparison, keeping it only in the Supplement for completeness.

RC3: The NSE is never formally defined. Please justify the choice of this metric as Schaeffli2007 have shown that NSE values can be inflated in highly seasonal regimes such as glacial and nival regimes.

AR: We will define NSE and KGE in Sect. 3.6 and add a Discussion paragraph on benchmark dependence in seasonal regimes, citing Schaeffli and Gupta (2007); RC2 raised the same point. NSE is retained as the primary metric because it is the one used by the studies we compare against, and we will report KGE alongside it. We may add a seasonal-benchmark efficiency for Experiment 1 in the Supplement.

RC3: *L9: when and how?*

AR: The abstract sentence will be made specific (which datasets, which period, which training pools).

RC3: *L17-18: unclear sentence.*

AR: Will be rewritten.

RC3: *L100: citation is missing, please check other citations as well.*

AR: Corrected; a full citation audit will be performed (also requested by RC1 and RC2).

RC3: *Fig 1 caption: North American glacierized catchments.*

AR: Accepted.

RC3: *L274-294: explain more intuitively for readers unfamiliar with GNNs.*

AR: We will rewrite the Graph WaveNet description in plain terms (each basin is a node; edges connect basins with similar static attributes; the network passes information between connected basins while a temporal convolution handles the time series), and move the literature background to the Introduction as RC1 requested.

RC3: *3.2: Consistently mention which example catchments are shown in which supplementary figure.*

AR: Accepted; the example basins and their supplementary figure numbers will be listed once in Sect. 3.

RC3: *L786: future work. L271: check sentence. L623: missing point.*

AR: All accepted.

Summary of changes in response to RC3

- Scope: Experiment 4 to the Supplement, Experiment 3 retained; full-period NSE retained, corrected melt-season analysis (Sect. 4.2, Table 5, Figs. 5 and 7) to the Supplement; Fig. 4 to the Supplement.
- Length: Introduction and Discussion condensed; Fig. 2 redrawn; Tables 2, 3 and 6 and the computation-time material to an appendix.
- Glacier classes: low (<2 %, 188 basins) and glacier-influenced (≥ 2 %, 95 basins) throughout; glacier-fraction histogram added; definition of G stated; minimum-cover definition recommended.
- Temporal split: stated as a limitation with Garreaud et al. (2025) and Hogan and Lundquist (2024); period boundaries explained and reconciled between Sect. 2.1.1 and Sect. 3.6; temporal re-sampling named as future work.
- Glacier-runoff input: disaggregation compared with Wiersma et al. (2022); absence of validation data and use of catchment-mean temperature stated.
- Inputs: ERA5-Land SWE justified with its known biases and with the common-input re-run of Experiment 1, reported in the Supplement (AC1, G1); glacier fraction stated not to be a model input.
- Metrics: NSE and KGE defined; benchmark dependence of NSE discussed; agreement between metrics judged on KGE and pBias.
- Data description: CAMELS-DK statement in Sect. 2.1.1 corrected; criterion for M stated.
- Technical corrections: captions of Figs. 1, 3–6; L9, L17–L18, L100, L271, L274–L294, L623, L786; citation audit.

References

- Garreaud, R., Boisier, J. P., Alvarez-Garreton, C., Christie, D. A., Carrasco-Escaff, T., Vergara, I., Chávez, R. O., Aldunce, P., Camus, P., Suazo-Álvarez, M., Masiokas, M., Castro, G., Muñoz, A., Zambrano-Bigiarini, M., Fuster, R., and Godoy, L.: Hyperdroughts in central Chile: drivers, impacts, and projections, *Hydrology and Earth System Sciences*, 29, 5347–5369, <https://doi.org/10.5194/hess-29-5347-2025>, 2025.
- Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Cohen, D., and Gilon, O.: How to deal w ___ missing input data, *Hydrology and Earth System Sciences*, 29, 6221–6235, <https://doi.org/10.5194/hess-29-6221-2025>, 2025.
- Hogan, D. and Lundquist, J. D.: Recent upper Colorado River streamflow declines driven by loss of spring precipitation, *Geophysical Research Letters*, 51, <https://doi.org/10.1029/2024GL109826>, 2024.
- Karagulle, D., Frye, C., Sayre, R., Breyer, S., Aniello, P., Vaughan, R., and Wright, D.: Modeling global Hammond landform regions from 250-m elevation data, *Transactions in GIS*, 21, 1040–1060, <https://doi.org/10.1111/tgis.12265>, 2017.
- 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.
- Lundquist, J. D., Abel, M., Currier, W. R., and Jackson, D. L.: Snow what? Comparing snow products to reveal what works, where, and why, *ESS Open Archive* [preprint], <https://doi.org/10.22541/essoar.177100538.80471917/v1>, 2026.
- Mudryk, L., Mortimer, C., Derksen, C., Elias Chereque, A., and Kushner, P.: Benchmarking of snow water equivalent (SWE) products based on outcomes of the SnowPEX+ Intercomparison Project, *The Cryosphere*, 19, 201–218, <https://doi.org/10.5194/tc-19-201-2025>, 2025.
- Schaefli, B. and Gupta, H. V.: Do Nash values have value?, *Hydrological Processes*, 21, 2075–2080, <https://doi.org/10.1002/hyp.6825>, 2007.

Wiersma, P., Aerts, J., Zekollari, H., Hrachowitz, M., Drost, N., Huss, M., Sutanudjaja, E. H., and Hut, R.:
Coupling a global glacier model to a global hydrological model prevents underestimation of glacier runoff,
Hydrology and Earth System Sciences, 26, 5971–5986, <https://doi.org/10.5194/hess-26-5971-2022>, 2022.