

Author Comment (AC1) in response to Referee Comment 1 (RC1) 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: Iiro Seppä (RC1). Final author comment, 2 October 2026.

We thank the referee for a careful and constructive review. The referee's comments have led us to sharpen the central claim of the paper and substantially streamline the manuscript. Below we respond to the seven general comments first (Part 1), collecting the specific comments that are clearly related to each general comment under that response, as the referee kindly allowed. All remaining specific comments are answered individually in Part 2. Technical corrections we simply accept are listed in Part 3. A summary of all planned changes is given in Part 4.

Part 1 — General comments

G1. Different meteorological forcings for δ HBV versus the DL models

RC1: *The main issue I have is that the three models are forced with different meteorological inputs, which invalidates any direct comparison between δ HBV and the two other models. Since δ HBV is the most limited in inputs it can receive, the DL models should be forced with those inputs, and nothing else. Here are two suggestions on how to possibly fix the issue: The best option is to rerun the DL experiments with same forcings. This is of course computationally expensive. Second option is to only compare the models with same inputs. This would be technically easy, but limits the usefulness of results and requires very careful and clear writing and plot organization, and lots of warnings that the models cannot be compared. I would only choose this if rerunning the models is unattainable due to computational limitations.*

Related specific comments collected here: L178, L180, L668, L679, L717–L719, L727.

AR: We agree with the referee, and we thank them for catching a statement in our own manuscript that is contradicted by our own supplementary tables. As Tables S2a and S2b show, the LSTM and Graph WaveNet receive five ERA5-Land dynamic inputs while δ HBV receives the three that the HBV structure admits (precipitation, temperature and potential evapotranspiration). The sentence at L180 ("All three models were forced with the same ERA5-Land meteorological fields... do not confound comparisons across architectures") is therefore incorrect as written, and we will remove it. We also accept the referee's point (L717–L719) that the original reason given in the Discussion — that δ HBV would have had to be rewritten — does not answer the question, because the DL models could simply have been restricted to the same three inputs. Our design choice was to give each architecture the input set that a practitioner would use "off the shelf" (Kratzert et al., 2019 and Sun et al., 2021 for the DL models; Feng et al., 2022 for δ HBV), but we accept that this makes the comparison one of deployed configurations, not of architectures under controlled inputs.

We followed the referee's first option in a targeted form. A full re-run of the experimental campaign (3 architectures $\times$ 5 experiments $\times$ up to 100 trials) with a common input set was not attainable within the final-response period, so we re-ran the experiment on which the architecture comparison rests and fixed in advance the rule for whether the remaining experiments would also need to be re-run: if restricting the DL models to the δ HBV inputs changed the ranking among architectures, or changed the partition effect (A vs. M vs. G) beyond the trial-to-trial variability, the common-input re-run would be extended to the remainder of experiments; otherwise it would be reported as a forcing-sensitivity check and the original experiments kept as the main results.

Stage 1 (completed). We retrained the LSTM and Graph WaveNet on the δ HBV input set (ERA5-Land precipitation, temperature and potential evapotranspiration) for Experiment 1, covering all three training partitions (A, M, G) with 10 trials per configuration. This input set excludes surface net solar radiation, snow water equivalent and surface pressure, which the DL models received in the original runs. δ HBV did not need to be re-run, as it already uses this input set. Figure R1 compares the test-basin NSE distributions of the original runs (Baseline) with those of the re-run (δ HBV -inputs). Ten trials are sufficient to detect a systematic shift in performance or in the ranking among architectures; they do not reproduce the full trial-to-trial spread of the main experiments.

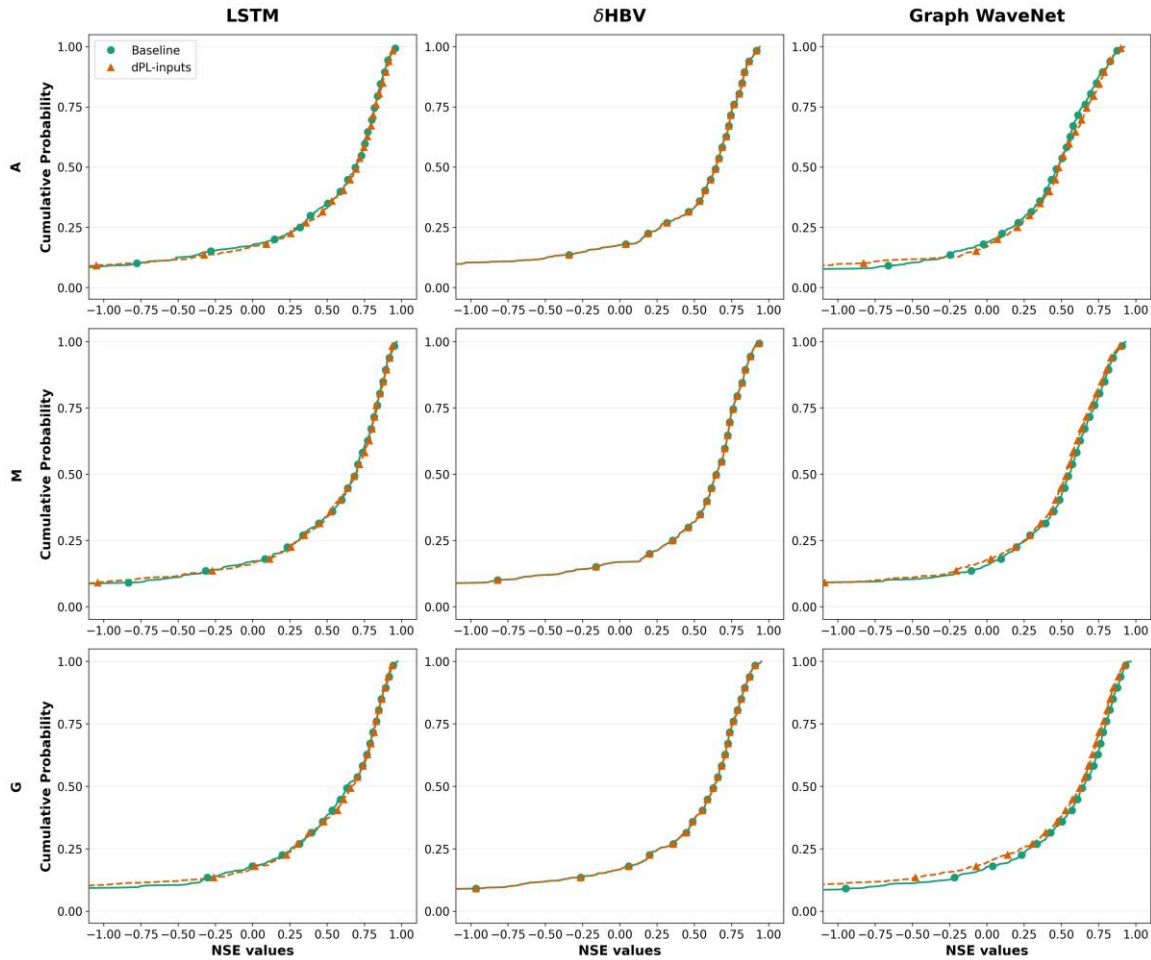


Figure R1. Cumulative distributions of test-basin NSE in Experiment 1 for models trained on partitions A, M and G (rows). Baseline: the original input sets (five ERA5-Land inputs for the LSTM and Graph WaveNet, three for δ HBV). dPL-inputs: the LSTM and Graph WaveNet retrained on the three δ HBV inputs. δ HBV was not re-run, so its two curves coincide and are shown for reference. The x-axis is truncated at NSE = -1.

Results. Restricting the DL models to the δ HBV inputs changes their performance very little (Fig. R1):

- LSTM: the NSE distributions with the two input sets are nearly indistinguishable in all three partitions, including the lower tail.
- Graph WaveNet: the differences are small and not consistent in sign. With the reduced inputs, NSE is slightly higher over most of the distribution in A and slightly lower in M and G (by a few hundredths of NSE near the median), and the lower tail is somewhat heavier in A and G, i.e. slightly more test basins have strongly negative NSE.
- Across architectures: the differences among the three models are the same as with the original inputs (the LSTM and δ HBV ahead of Graph WaveNet in A and M, and the three close together in G), and the direction of the partition effect within each model is unchanged.

What this means for the conclusions. By the rule stated above, Stage 2 is not triggered, and the original experiments remain the main results. Two conclusions follow. First, the primary finding of the paper, which compares each architecture with itself across training partitions, holds when all three models receive the same inputs. Second, in Experiment 1 the differences between δ HBV and the DL models are not explained by the additional inputs the DL models received: the LSTM performs the same without them, and Graph WaveNet changes only slightly. The additional inputs, including

the ERA5-Land snow water equivalent and net solar radiation questioned by RC3 and RC2, therefore do not drive the results in this setting. Experiments 2–5 were not re-run, so for those experiments the comparison across architectures remains one between the configurations as deployed.

How the manuscript will be revised. The following changes will be made:

- The Stage 1 re-run will be described in Sect. 3 and reported in the Supplement (Fig. R1, with a table of median NSE and interquartile range per model, partition and input set), and referenced wherever the architectures are compared.
- The sentence at L180 will be removed, and Table 2 will gain a column listing the dynamic inputs of each model, so that the difference is visible in the main text.
- The fact that the primary finding rests on within-architecture comparisons will be made explicit in the Introduction, at the end of Sect. 2.1.1 and in the Discussion, citing the re-run as confirmation under common inputs.
- Statements comparing δ HBV with the LSTM or Graph WaveNet (Abstract; Sect. 4; L668; Conclusions) will state that in Experiment 1 the comparison holds under common inputs, and will describe the comparisons in Experiments 2–5 as comparisons between the configurations as deployed. The sentence at L727 will be deleted, and the explanation at L717–L719 replaced by a description of the re-run.
- The limitations paragraph in Sect. 5 will state the input difference, that the common-input check covers Experiment 1 with 10 trials, and that Experiments 2–5 were not re-run with common inputs.

G2. Missing references

RC1: *A worrying amount of citations where I wanted to check the original source did not have a corresponding reference. Please carefully check all the citations and ensure that they have a reference.*

Related specific comments collected here: L82, L92, L100, L108, L111 (missing), L198 (missing), L206–L207, L22, L23, L146.

AR: We apologise for this. A late-stage change of reference-management software introduced a mismatch between in-text citations and the reference list. We will run a full citation audit of the manuscript and supplement (every in-text citation checked against the list and vice versa).

G3. Section "spill-over"

RC1: *Please ensure that the sections don't 'spill over' into other sections. For example, there were part that belonged to methods in the results, new results in the discussion and introduction in methodology. End of introduction is a bit freer, but other than that the structure should be the following: 1. introducing the important literature and explaining what is the research gap. 2. Data and methods (methods should only contain things you did, often with brief explanation why with a citation, but no other literature review). 3. results (don't include data description here, it goes to section 2.) 4. discussion. Don't introduce new results here (computation time).*

Related specific comments collected here: L254–L264, L274–L279, L295–L310, L318–L319, L363–L364, L397, L512, L539, L627–L628, L680, L683–L684.

AR: We agree and will restructure along the lines the referee sets out. Specifically:

- Literature and motivation currently in the Methods (L254–L264 on LSTM background, L274–L279 on GNNs, L295–L310 on differentiable/hybrid modelling) will be condensed and merged into the Introduction, removing overlap with material already there. The Methods

will retain only what we did, with citations. When merging L295–L310 we will add the point raised by the referee that the LSTM inside a hybrid model can learn around the conceptual constraints (Acuña Espinoza et al., 2024; Álvarez Chaves et al., 2026).

- The introductory paragraph of Sect. 4.3 (L512) and the background material in Sect. 4.3.2 (L539) will be moved to Sect. 2.1.3 (glacier inputs); only results will remain in Sect. 4.
- The computation-time material (L680, L683–L684, Table 6) will be removed from the Discussion. We agree with the referee that it is a result rather than discussion, and that it is useful mainly for model selection rather than for the question the paper addresses. It will therefore be placed in an appendix, as the referee suggests, where it remains part of the manuscript and can be cited as Table A1, with a one-sentence pointer from the Discussion.
- Redundant statements will be deleted (L318–L319, already at L311; L363–L364, redundant with the previous section) and the training/testing overview at L397 will be moved to the start of Sect. 3 so that it precedes the experiment descriptions and is stated once. The background at L627–L628 will be moved to the Introduction or deleted if already present.

G4. "Regional" versus "climatic" generalisation

RC1: *This study talks a lot about regional generalization. However, it only studied climatic generalization, so any regional claims should be limited to situations where you are citing literature that actually studied regional generalization.*

Related specific comments collected here: L111, L114 (terminology), L121, L150, L152, L153, L648, L787, L788–L789.

AR: We accept this. Our experiments hold out basins by hydroclimatic class (glacierized / mountain / other), the region containing a held-out basin could remain represented in the training pool, so the design tests prediction in ungauged basins and, across the A → M → G gradient, prediction in a hydroclimatic class that is under- or un-represented in training. It does not test regional transfer: we did not, for example, train on Alpine basins and test on the Rocky Mountains. We will:

- (i) adopt the terminology proposed by the referee — prediction in ungauged basins (PUB) for any held-out basin, prediction in ungauged regions (PUR) for regional hold-outs, and prediction in ungauged climates (PUC) for hold-outs by hydroclimatic class — defining it at first use in the Introduction, noting that we introduce PUC for hold-outs by hydroclimatic class, and apply it consistently in the title of Sect. 3, the figure captions and the Discussion;
- (ii) describe our experiments as PUC (with PUB in space and time as the baseline) and apply this consistently throughout;
- (iii) replace every occurrence of "regional generalization", "data-scarce regions", "out of sample in landscape" and similar with "hydroclimatic generalization" / "generalization to a hydroclimatic class not represented in training";
- (iv) delete the clauses at L648 and L787–L789 that make regional-transfer claims.

Where we cite literature that did test regional transfer (e.g. Feng et al., 2024 on PUR) we will say so explicitly and keep it separate from our own claims. We note for completeness that exclusion by hydroclimatic class is not fully independent of geography, since glacierized and high-mountain basins are geographically clustered; the exclusive experiments therefore reduce the geographic as well as the hydroclimatic coverage of the training pool near the target basins. Because region is not varied independently in our design, we draw no regional conclusion from this and confine ourselves to the PUC framing throughout.

We will retain the title, which is accurate under this terminology: the test basins are ungauged in the PUB sense, and the sub-title describes the origin of the data rather than asserting transfer between regions.

G5. Experiment 5 is separated from the other experiments and over-weighted

RC1: *The methods and results should be thoroughly revised to present the fifth experiment along with the rest. The fifth experiment seems currently completely separated from the rest without any good reason. In my opinion it has also been given too much space since its result could be summarized in a few sentences.*

Related specific comments collected here: L334, L349, L414, L430, L575, L621, L732–L733.

AR: We agree. In the revised manuscript Experiment 5 (G partition with and without daily glacier-runoff input) will be introduced in Sect. 3 in the same format as Experiments 1–4, its results will be added as a row/panel in Fig. 3 and Table 4 alongside the other experiments (as also requested at L430), and the present Sect. 4.3.2 with Figs. 8–9 will be reduced to a short paragraph. The pBias $\pm$ band analysis (L575) will be dropped: we agree that the band boundaries are arbitrary and that the main message — that the disaggregated glacier-runoff input yields at best a small change in skill — can be stated in a few sentences. In place of the band bookkeeping we will report the paired per-basin difference in NSE (with and without glacier input, same basins and same trials) with its distribution and a paired significance test, as suggested. The count inconsistencies ("four experiments" at L334, L349 and L621) will be corrected to five. We agree with the referee's reading at L732–L733 that the Experiment 5 result implies an explicit glacier module is not a prerequisite for skill in this setting, and we will reword that limitation accordingly.

G6. Focus: everything should support Fig. 3 / Table 4

RC1: *There are a lot of sections where the text feels like the article starts looking at something that is a diversion from the main focus. I think the most interesting part of the research can be summarized by figure 3 and table 5 (experiment 5 should be included in those), and I think that everything in the article should focus strongly to support that. Everything in the results or methods that is not relevant for the point of figure 3 should be moved to appendix or supplement in my opinion (for example performance considerations).*

Related specific comments collected here: L219–L221, L410, L418, L435, L479, L486, L515–L516, L525–L526, L535, L638.

AR: We agree, and this comment together with the specific comments has guided a substantial shortening. (We read "table 5" as Table 4, the full-period NSE table; Table 5 is the summer table addressed below.) Referees 2 and 3 additionally recommend reducing the number of experiments; following RC2, Experiment 4 (exclusive full-set selection) will be moved to the Supplement, which also serves the shortening this comment asks for. Experiment 5 will be integrated into the main text as described under G5, so the main text will present four experiments and Fig. 3 and Table 4 will carry all four.

In deciding between an appendix and the Supplement for each item, we have followed the principle that material a main-text conclusion rests on should remain part of the manuscript, in an appendix, while material that is complementary to the argument belongs in the Supplement. Specifically:

Moved to an appendix:

- Tables S5 and S6 (KGE, RMSE and pBias for all experiments) will be moved from the Supplement to an appendix, as the referee asks at L418, and the Results will state for each experiment whether the other metrics agree with NSE. Because the claim that our

conclusions are not an artefact of the choice of metric rests on these tables, they belong in the manuscript rather than in the Supplement.

- Computation time (Table 6) moves from the Discussion to the same appendix, as noted under G3.

Moved to the Supplement:

- Experiment 4 moves to the Supplement with its own short results paragraph; the main text will state once why it is reported there (it varies training volume under the same exclusivity as Experiment 3 and did not change the conclusions).
- The summer-months analysis (Sect. 4.2, Table 5, Fig. 5; L410, L479, L486, L525–L526, L535) will be removed from the main text. The referee is right on both counts: our June–August definition is wrong for the Patagonian basins in the Southern Hemisphere, and sub-period NSE is biased low by construction (Klotz et al., 2024), so the section cannot support a comparison as written. We will retain a corrected version in the Supplement (hemisphere-aware melt-season definition, RMSE instead of NSE) for readers interested in the seasonal behaviour, with one sentence in the Results pointing to it.
- Fig. 4 (L435), which duplicates Fig. 3 and in which the M and G curves are not visually separable, will be moved to the Supplement.
- The representative single-basin hydrographs (Sect. 4.3.1, L515–L516) will be moved to the Supplement; we agree one basin cannot represent a class and that the paper is about groups of basins. (Referee 2 asks that two example hydrographs be retained in the main text; we will keep two, labelled as examples rather than as representative basins, and move the remaining panels to the Supplement.)

Removed altogether:

- The RGI-region categorisation (L219–L221), which is not used in any analysis, will be removed; the geographic description will be reduced to one sentence in Sect. 2.1.1.

Also in response to this comment, study-specific abbreviations will be written out in the Discussion and Conclusions (L638) so that readers who jump there can follow.

G7. Specific comments in the PDF

RC1: *Specific comments and technical corrections are highlighted and commented to the attached pdf. If specific comments are clearly related to a general comment, I'm fine with one response collecting these comments together. No need to respond to technical corrections if you agree with them.*

AR: Thank you. Comments clearly related to a general comment are collected above; all remaining specific comments are answered in Part 2, and the technical corrections we accept without further comment are listed in Part 3.

Part 2 — Specific comments answered individually

Comments already collected under a general comment above are listed here only with a pointer.

RC1: L34: *what is the difference between glacier and glacier-ice here?*

AR: None is intended. We will write "snow and glacier ice" once and remove the redundant term.

RC1: L35: *you are talking about snow earlier in the sentence, and now discussing the time when it has already melted. Please modify to make the sentence sensical.*

AR: We will split the sentence: one on the seasonal snowpack and its melt pulse, a second on glacier ice sustaining late-summer flow after the snowpack is gone.

RC1: L57-L59: *Each point should cite an example original study or a few, citing the review is not enough here. The citations should follow immediately after their point.*

AR: Agreed. Each of the listed challenges will be followed immediately by one or two primary studies rather than the review alone.

RC1: L82 / L92 / L100 / L108 / L111: *missing or duplicated references.*

AR: See Part 1 G2.

RC1: L111: *Please explain why this is an limitation. Many studies have the explicit goal of generalizing to a specific region ... If you mean exposure to specific climates and the effects of performance of that, then use that term. Citations are also needed ... Nearing et al. 2024 could be an interesting reference.*

AR: We agree that the sentence mis-states the issue, and we accept that generalising to a specific region is a legitimate objective rather than a shortcoming. What we intended to identify as a limitation is narrower: where training and evaluation draw on the same region — the "semi-gauged" setting of Gleason and Durand (2020) — the resulting skill does not tell us how the model behaves when the target hydroclimate itself is under-represented in the training pool. That is the situation for glacierized basins in the large-sample datasets used to train global DL models: in our own pool, 283 of 2,845 basins (about 10 %) are glacierized, so a model trained on the full pool sees comparatively few of them. Using the terminology adopted under G4, the open question is one of prediction in ungauged climates (PUC) rather than of regional performance. We will rewrite the sentence accordingly, and cite Nearing et al. (2024), whose global model is applied in regions with sparse gauge coverage.

RC1: L113: *'comparison of training data conditions' is vague and uninformative ... Please explain what it should be compared against ... why this is relevant for the selection of the architecture.*

AR: We will rewrite the sentence to state concretely what is compared (training pools of different size and different hydroclimatic representativeness relative to the target class), against what baseline (the same architecture trained on the largest pool), and why it bears on architecture choice (architectures may differ in how much they depend on representativeness versus volume), with citations to Gauch et al. (2021) and Kratzert et al. (2019).

RC1: L114: *I strongly suggest deciding (and explicitly defining) a clear terminology, such as PUB / PUR / PUC ...*

AR: Adopted; Part 1 see G4.

RC1: L114: *condition sounds like a quality issue, while this sentence makes it feel like (spatial) distribution would be a more proper word.*

AR: Agreed that "conditions" wrongly suggests data quality. We will use "training-data composition" — size and hydroclimatic coverage, as defined at L113 — in place of all five occurrences. We prefer

this to "spatial distribution" because what varies across the A, M and G partitions is hydroclimatic coverage rather than geography; "spatial" would reintroduce the regional framing removed under G4. The term is already used in this sense in the Discussion of the submitted manuscript.

RC1: L118: *This statement is currently misleading, because there are also examples that include glacial regions, such as Kraft et al. (2025) and Aguayo et al. (2025). These have a lot of relevance for this article, and should be integrated also to discussion.*

AR: Thank you for these references. We will correct the statement, cite both studies in the Introduction as examples of DL/hybrid models applied to glacierized regions, and add a paragraph to the Discussion relating our Experiment 5 result (little gain from an externally modelled glacier-runoff input) to the hybrid glacio-hydrological approach of Aguayo et al. (2025) and the spatially contiguous reconstruction of Kraft et al. (2025).

RC1: L119: *I think it would be beneficial to use clear terminology that separates the original CAMELS dataset for CONUS from the family of CAMELS datasets.*

AR: We will write "CAMELS-US" for the original CONUS dataset and "the CAMELS family of datasets" (with members named) for the group, defined at first mention and used consistently.

RC1: L119: *... CAMELS (USA or elsewhere) are not global datasets, so the sentence feels confused. CARAVAN also includes glaciated catchments. There are CAMELS datasets that do include glaciers ... CAMELS-CL, CAMELS-SE, CAMELS-CH, LAMAH-CE, LAMAH-ICE. If you do have some examples of truly global datasets that for some reason exclude cryospheric catchments, please cite them.*

AR: We agree the sentence conflates two things and will rewrite it. The point we intend is different and narrower: glacierized basins are a small minority of the basins in the large-sample datasets used to train global DL models. Table S1 of the submitted Supplement already lists our pool by dataset with glacierized-basin counts — 283 of 2,845 basins, about 10 % — including CAMELS-CL, CAMELS-CH and LamaH-CE, which the referee correctly notes do contain glaciers.

RC1: L121: *... This claim is too general. For example, Feng et al. (2024) show that LSTM outperforms δ HBV in PUB, but the relation is the other way around for PUR. PUB results of Feng et al. (2022) also depended on the meteorological forcing used ...*

AR: We will replace the general claim with the differentiated statement the referee gives: LSTM tends to outperform δ HBV in PUB but δ HBV in PUR (Feng et al., 2024), and PUB rankings depend on the forcing product (Feng et al., 2022; Kratzert et al., 2021). This also motivates our G1 caveat that forcing choice can drive apparent architecture differences.

RC1: L122-L123: *This sentence does not work currently ... Liu et al 2025a were not global, multforcing or doing anything PUB-related ... Song et al. 2026 was also not global, and there LSTM was outperformed only when looking at the rarer events ... (Baste et al. 2025).*

AR: We will delete the sentence. Liu et al. (2025a) will be cited only where its actual scope applies (or removed), and Song et al. (2026) will be re-cited in the Discussion in the context the referee describes — LSTM under-performance at rare high flows, linked to the extrapolation limits documented by Baste et al. (2025), which we will add.

RC1: L123: *please specify which Feng et al. 2024; ... mark it clearly with a and b if it is a separate study.*

AR: See Part 1 G2.

RC1: L125-L129: *This sentence is not very readable, and should probably be split into multiple sentences. L131-L132: I think this part would be more readable if moved after the main sentence.*

AR: Both accepted.

RC1: L146: *This Liu is not the same person as Liu 2025a ...*

AR: Will be distinguished by initials; see Part 1 G2.

RC1: L147: *... your graph for the Graph WaveNet was based on attribute similarity instead of physical spatial structure ... in geosciences the term [spatial] should only be used when you can use meters as the unit.*

AR: We agree. The graph is built from static-attribute similarity, not from geographic distance or river-network topology. We will state this at first mention and relabel the paradigm "graph-based (attribute-similarity)" rather than "spatial" throughout, including in Table 2 and the figure legends.

RC1: L149-L150: *... the sentence should maybe be split into multiple sentences.*

AR: Accepted.

RC1: L150 / L152 / L153: *terminology; not data-scarce regions but excluded climates; "target catchment types".*

AR: See G4. We will adopt "target hydroclimatic class" (glacierized high-mountain catchments) in place of "target region".

RC1: L164: *CARAVAN has seen a major expansion in version 1.3 ... Please state your used version in the article and explain in your response why you did not use that ...*

AR: We used Caravan v1.0 (Zenodo 7540792), with the CAMELS-CH extension. The campaign for Experiments 1–5 was designed and largely executed before the GRDC-Caravan extension (v1.3; Färber et al., 2025) was available, and re-running it on the enlarged dataset exceeds our computational budget. We agree the additional glacierized basins from Norway, New Zealand, Argentina and Iceland would improve diversity. We will state the version in Sect. 2.1, cite Färber et al. (2025), and recommend v1.3 for future work in the limitations, noting that whether the results hold on the larger sample remains to be tested.

RC1: L172: *CAMELS-DK doesn't include any glaciatic catchments, so the sentence should be modified to only point to CAMELS-CH.*

AR: Correct; we will fix this and explain that CAMELS-DK was included only to enlarge the A partition.

RC1: L174: *how continuous? Were any small gaps allowed? If yes, how long?*

AR: We will add this to Sect. 2.1.1.

RC1: L178: *Why are there separate forcings for the DL and the hybrid model? ... BTW, in table S2a the unit of snow water equivalent should be mm, not (mm/day).*

AR: See G1 for the forcing question. The SWE unit in Table S2a will be corrected to mm.

RC1: L180: *If the supplement information is correct, you used different features for the models, and then it doesn't matter if they all come from ERA5.*

AR: The supplement is correct and the sentence at L180 is wrong; it will be removed. See G1.

RC1: L185-L186: *What is meant by this? The amount of glacier catchments doesn't increase with the sort of grouping done.*

AR: The sentence is misleading and will be rewritten. What increases across $A \rightarrow M \rightarrow G$ is the proportion of glacierized basins in the training pool (i.e. representativeness), not their number, which is the same 283 in all three.

RC1: L198-L200: *This is redundant with the previous sentence ...*

AR: Accepted.

RC1: L198: *Missing from the references. If this is Karagulle et al. (2017) ... are there any relevant glaciers in the data that are outside of the high mountains, since the classification is purely topography based?*

AR: Yes, it is the Hammond landform classification of Karagulle et al. (2017); the reference will be added. The referee is right that the classification is topographic. Because our G partition is defined as the intersection of the high-mountain landform class with RGI glacier outlines, glacierized basins that lie outside the high-mountain class — plausibly low-elevation outlets in southern Chile — are excluded from G by construction. We will state this explicitly in Sect. 2.1.1 as a definitional limitation.

RC1: L200: *Link leads to 404 error.*

AR: Will be replaced with a working URL/DOI.

RC1: L206-L207: *Missing from the references.*

AR: See Part 1 G2.

RC1: L213: *Why is CAMELS-DK caravan extension missing from here ... why are there so few catchments from Chile? ... wouldn't it have been worth it to select slightly shorter period in order to get more catchments.*

AR: CAMELS-DK is not part of our training pool — Table S1 lists the datasets actually used, summing to 2,845 basins. We used only the CAMELS-CH extension. CAMELS-DK contains no basins meeting our mountain or glacierized criteria, so its absence affects only the size of the A partition and not the M or G partitions on which our comparisons rest. We will state this in Sect. 2.1.

The small Chilean sample follows from the continuous-record requirement over the training period 1988–2008. We agree a shorter period would have admitted more and will note the trade-off in the limitations; re-running on a different period is beyond our budget.

RC1: L214: *Actually, there are now catchments from (Southern) Africa.*

AR: We will correct the statement to refer to Caravan v1.0 as used here and note that v1.3 (Färber et al., 2025) adds African basins.

RC1: L219-L221: *this categorization is not used for any actual analysis ... should be removed.*

AR: See Part 1 G6; removed.

RC1: L222: *Why is this exact set of values used? ... high and very high classes have too few observations ...*

AR: We agree the high (4 basins) and very-high (2 basins) classes are too small to support comparison, and, as the referee notes at Sect. 4.3.1, our analysis in practice contrasts low-glacier basins with the rest. We will therefore adopt two classes throughout the analysis — low (<2 %, 188 basins) and glacier-influenced (≥2 %, 95 basins) — stating counts in Table 1 and in every figure caption. The finer bins will be retained in Table 1 and the Supplement as a description of the sample, with no comparative claims drawn from them.

RC1: L232: *Is the temperature based on the areally averaged temperature of the catchment? ... mean areal temperature of the catchment may not be representative of the conditions on the glacier.*

AR: Yes — the positive-degree-day weights use the catchment-mean ERA5-Land temperature from Caravan, and we will state this explicitly in Sect. 2. We accept the concern: catchment-mean temperature is not representative of glacier-surface conditions, and the mismatch varies between basins, being largest where the glacierized area lies well above the catchment mean elevation. We will add this as a limitation and note that an elevation-band or lapse-rate correction would be the

appropriate refinement. Its effect is bounded to how each month's runoff is distributed across days, since the monthly volume comes from PyGEM using glacier-specific climate.

RC1: L237: *I think the notation is not as clear as it should be ... m and j are completely unexplained ... state if you are using vectors, and also their shape.*

AR: We will rewrite the equations with every symbol defined.

RC1: L245-L246: *This reference is quite old, are you sure that there aren't any newer and more sophisticated methods?*

AR: More sophisticated daily glacier-runoff models certainly exist (e.g. energy-balance approaches), but our aim was only to distribute a monthly total, for which the degree-day weighting is the standard minimal choice.

RC1: L254-L264 / L274-L279 / L295-L310: *move to introduction ...*

AR: See Part 1 G3.

RC1: L271: *Why this exact set of hyperparameters? ... How about the values of other hyperparameters, such as hidden size? Did you use dropout or add initial bias to the forget gate ...*

AR: No hyperparameter tuning was performed. The LSTM configuration follows Kratzert et al. (2019) as implemented in NeuralHydrology, the Graph WaveNet configuration follows Sun et al. (2021), and δ HBV follows Feng et al. (2022). We will state this and add a table to the Supplement listing all hyperparameters (hidden size, dropout, learning-rate schedule, forget-gate bias initialisation, TCN/GCN layer settings) for each model

RC1: L318-L319: *Remove this, it is stated already at line 311.*

AR: Accepted.

RC1: L344: *space and time.*

AR: Accepted.

RC1: L352: *It would probably be good idea to state somewhere around here that the training set is also redrawn hundred times.*

AR: Accepted.

RC1: L353: *This setup can still produce biases ... Since the source catchments are heavily concentrated to The Alps and North America, there is a strong regional bias ... please be more specific what sort of bias you are looking to reduce.*

AR: Agreed. We will replace "not biased" with a precise statement: the 100 draws remove sensitivity to any single choice of test basins (sampling variability), not the regional composition of the G partition, which they preserve by construction. We will state that composition explicitly — of 283 G basins, 130 are Alpine, 147 North American and 6 Chilean (Table S1) — so pooled results are weighted towards the first two, and we will scope our conclusions to that sample rather than to glacierized basins generally.

RC1: L363-L364: *Redundant with previous section.*

AR: Deleted; see G3.

RC1: L366: *The cited works are experimental, not theoretical, so 'previous ML experiments' would be more appropriate.*

AR: Accepted.

RC1: L369-L370: *It would be interesting to see a comparison of the other models at 36 replicates ... Supplements would probably be the best place for that.*

AR: We will add a supplementary figure comparing all three models on the first 36 trials only, alongside the full-trial results, to show whether the truncated ensemble differs from the full one.

RC1: L374: *I think it would be enough to state this just once at Line 351. Additionally, are different experiments using the exact same test set or is it randomized again for different experiments?*

AR: For a given trial the test set is identical across the three models, as stated at L351. It is not identical across experiments: each experiment draws its own 100 test sets from the G pool. Comparisons across experiments are therefore between distributions over 100 draws rather than between the same basins, which is why we report distributions rather than paired per-basin differences across experiments; within an experiment, model comparisons are paired. We will state this once at L351 and remove the repetition at L374.

RC1: L388-L389: *... this could be written clearer. Is this trial size limitation affecting only training, or also testing? ... testing is cheap so if you have not run the testing for the full set, please do so.*

AR: We accept this; the referee is right that testing can be extended. In Experiment 4 the A and M training pools contain no glacierized basins and are therefore identical in every trial; the 10 trials differ only in random initialization. None of these 10 models has seen any G basin, so each can be evaluated out of sample on all 100 test sets at negligible cost, and we will do so and report the results in the revised (supplementary) Experiment 4. For the G partition, the training set is the complement of each trial's test basins, so a model from one trial has seen most of the test basins of the other trials; testing it on them would not be out of sample, and additional G trials require additional training. We will state this distinction at L388–L389.

RC1: L392: *why only the first 10?*

AR: Because Experiment 5 requires retraining all three models with an additional input, we limited it to 10 trials for computational cost. We will state this, and note that the 10 trials use the same test sets as trials 1–10 of Experiment 1, so the comparison is paired.

RC1: L395: *... are there any studies that recommend this sort of practice [adding glacier runoff to precipitation for δHBV] or study its effects?*

AR: We are not aware of a study that evaluates this practice; it is a pragmatic workaround because δHBV has no input port for an external runoff term. We will say so, and we will state explicitly (see also L665) that adding glacier runoff to precipitation double-counts water in the model's mass balance and routes it through the soil and response stores as if it were rain — a plausible reason for the weak effect in δHBV .

RC1: L395: *please be consistent in the abbreviation you are using.*

AR: We will use " δHBV " throughout.

RC1: L396: *I think the figures here ... look like very quickly done placeholders ... upgrading these to more polished, higher quality figures.*

AR: Agreed. Figs. 2 and 3 (and the figures moved to the Supplement) will be redrawn at publication quality with consistent fonts, colours and axis formatting. Fig. 3 will also gain finer x-axis ticks and vertical gridlines as suggested at L430.

RC1: L397: *Consider moving this to before discussing the different experiments ...*

AR: Accepted; see Part 1 G3.

RC1: L410 / L479 / L525-L526 / L535: *summer definition (July–August) is wrong for the southern hemisphere ...*

AR: The referee is right; see G6 (Sect. 4.2 removed from the main text; corrected hemisphere-aware version in the Supplement).

RC1: L414: *I see no reason to separate the experiment 5 from the other experiments ...*

AR: See Part 1 G5.

RC1: L418: *... mention whether the other metrics generally agree with NSE ... Reference also the supplement tables s5 and s6 (those should be moved to appendix).*

AR: Accepted; see Part 1 G6.

RC1: L419 / L448 / L456 / L486: *Why 0.5 as the threshold for strong performance? ... I would recommend a more nuanced analysis of the values instead.*

AR: We accept this. We will remove the 0.5 "strong performance" threshold and the good/bad framing from the text, tables and figures, and describe the distributions instead (median, interquartile range and the full CDF already shown in Fig. 3). We will keep only the $NSE < 0$ threshold, for the reason the referee gives at L454.

RC1: L422-L424: *I disagree with this conclusion. There is a small improvement in LSTM performance, especially when adding other mountainous catchments ... I think a statistical test ... would be good.*

AR: We agree that "partition-agnostic" overstates the result and will soften the wording. Because the 100 trials share basins and are not independent, a test that pools NSE values across trials would overstate significance (Sect. 3.6). Within each trial, however, the test basins are identical across partitions, so the comparison can be paired. For each trial we will compute the per-basin difference in NSE between partitions on the same test basins, and report the distribution of the median paired difference across trials together with a paired test and an effect size. If the gain of the LSTM (and δHBV) from adding M basins is consistent in sign across trials, we will describe it as small but systematic; otherwise we will state that there is no consistent difference. The same analysis will be applied at L631 and to Experiment 5 (L575).

RC1: L430: *experiment five should be added here. Also, ... vertical lines along the x-axis ticks ... x-axis ticks a bit more frequently.*

AR: All three accepted (Fig. 3).

RC1: L435: *I think this figure is redundant with Figure 3 ...*

AR: Fig. 4 moved to the Supplement; see G6.

RC1: L454: *... using zero as a threshold for NSE is actually reasonable ... I think this point could be made explicitly ...*

AR: We will add the sentence that $NSE < 0$ means the model is worse than predicting the basin's mean discharge and is therefore a failure for essentially any purpose.

RC1: L462: *Shouldn't median NSE be a single value instead of a range?*

AR: The bracketed values are ranges across the three models, not ranges of a single model's median. We will rewrite the sentence to report the trial-median NSE per model rather than as a range — e.g. "median NSE under partition A is x (LSTM), y (δHBV) and z (Graph WaveNet)" — and do the same for partition M. We will also make clear that the improvement is relative to Experiment 3 and that absolute skill under partition A remains poor, with at least one model below zero. We will check other reported ranges in the Results for the same ambiguity.

RC1: L486: *NSE (or KGE) cannot be used robustly in this type of comparison ... (Klotz et al., 2024) ... I suggest either moving to use RMSE or removing this part entirely ...*

AR: See G6.

RC1: L512: *This section (4.3, not the subsections) should be in methods.*

AR: Accepted; see G3.

RC1: L515-L516: *How was the representativeness ensured? ... Consider if this is even necessary in the main body ...*

AR: Moved to the Supplement and reworded as "example" rather than "representative" basins; see G6.

RC1: L515: *Why are you talking about four-way division ... when you are actually presenting a reasonable division into two groups ...*

AR: Only the two-group split will remain; see L222.

RC1: L539: *Considerable portions of this section should be moved to section 2.1.3.*

AR: Accepted; see G3.

RC1: L575: *... present alongside the other experiments, and shortened considerably ... Just do a statistical test ...*

AR: See G5.

Section 5 — Discussion and Conclusions

RC1: L621: *first four, you had five experiments. L623: missing .*

AR: Both corrected.

RC1: L627-L628: *either to introduction or removed ...*

AR: See G3.

RC1: L631: *also experiment 1, because both LSTM and δ HBV improved (although only slightly ...)*

AR: Accepted; the paired analysis described at L422–L424 will be reported here for Experiment 1 as well.

RC1: L633: *I think that Gauch et al. (2021) should be integrated into this part of discussion.*

AR: Agreed; Gauch et al. (2021) will be discussed here.

RC1: L638: *abbreviations in the discussion. L639: a size limited.*

AR: Both accepted; see G6.

RC1: L648: *You did not do regional tests ...*

AR: Clause deleted; see G4.

RC1: L661: *I don't think it's that surprising or counterintuitive at all ... the data from Rounce et al. 2023 is being forced with ERA5 ... the issue more likely in the quality of the temporal signal ...*

AR: We accept this reading and will rewrite the paragraph. "Counterintuitive" will be removed. We will state that (i) the Rounce et al. (2023) runoff is itself driven by ERA5, so it is not independent of the other inputs; (ii) a daily product obtained from a monthly total by a simple temperature-index weighting carries little additional temporal information; and (iii) for δ HBV the precipitation workaround introduces its own distortion (L665). The Experiment 5 result is therefore expected rather than surprising, which is consistent with the referee's remark at L732–L733.

RC1: L663: *I think this sentence applies mostly to δ HBV, not to the other models.*

AR: Accepted; restricted to δ HBV.

RC1: L665: *I think could is not needed ...*

AR: Accepted; "could affect" → "affects".

RC1: L668: *when the training catchments are very different from the target catchments. Also, the models are currently not comparable due to different inputs.*

AR: Wording accepted; cross-architecture ranking removed per G1.

RC1: L670: *limited both in size and ... Otherwise this could be read that LSTM performed the worst when training with just G, which is not true.*

AR: Accepted; will read "limited both in size and in hydroclimatic representativeness".

RC1: L675: *Actually, the LSTM can learn to work around the deficiencies of the physical constraints ... it would be better to not make any strong claims about the constraints. An interesting point that you could make is that despite the above findings, in (extremely) target-climate poor setups the constraints seemed to lessen the reduction in performance.*

AR: We agree and will not make mechanistic claims about the constraints; we will not attempt the entropy-based analysis of Álvarez Chaves et al. (2026) in this paper. We will cite Acuña Espinoza et al. (2024) and Álvarez Chaves et al. (2026) to say that the constraints in δ HBV are known to be weak, and we will adopt the referee's descriptive framing.

RC1: L678: *representative would be better word. Your own experiments showed that LSTM was still the best or equal to GraphWaveNet when trained only with G.*

AR: Accepted.

RC1: L679: *You actually cannot compare the performance of δ HBV to the other two models ...*

AR: Sentence removed; see G1.

RC1: L680 / L683-684: *computation time is a result and belongs in an appendix ...*

AR: See G3.

RC1: L717-L719 / L727: *forcing comparison ...*

AR: See G1.

RC1: L732-L733: *The results from experiment 5 show in my opinion, that this might not be needed, so it is not really a limitation ...*

AR: Agreed; see G5.

RC1: L735: *This is outside the scope of what was tested ... If you know of a study that has shown that some glaciers will enter climates completely outside of current known climates please reference that.*

AR: We will remove the claim. We do not have a source showing glacierized basins entering climates absent from the historical record, and the referee's point that a basin losing glacier cover may simply come to resemble present-day basins with lower glacier fraction is well taken.

RC1: L740-L742: *This study doesn't actually have anything in the methods or results that would warrant such conclusions ... the scenario where we would have to try get a model performing well without any glacial catchment data is extremely unrealistic ... I still think exclusion of glacial catchments makes for an interesting experiment, but that doesn't mean the scenario is realistic.*

AR: We agree and will reframe. The glacier-exclusion experiments will be described as a controlled stress test of how strongly each architecture depends on representation of the target hydroclimatic class in training, not as a realistic data scenario; we will state that data from glacierized basins exist and are unlikely to disappear. The climate-change paragraph will be reduced to one cautious sentence: our results say nothing about the transferability of the models to future conditions, which would require a study designed for that purpose.

RC1: L746: ... *the methods of this study are too coarse to say anything concrete about the models ability to adapt to slightly different climate ...*

AR: Accepted; this sentence will be removed together with the change at L740–L742.

RC1: L766: *I would not call LSTMs behavior overfitting ... this is a matter of transferability.*

AR: Accepted; "overfitting" → "limited transferability to a hydroclimate absent from training".

RC1: L787: *I think this would equate to saying anywhere, which I disagree with. L788-L789: ... excluding catchments from a specific climate, not specific region.*

AR: Both accepted; the Conclusions will be rewritten under the PUC framing of G4 without regional claims.

Part 3 — Technical corrections accepted without further comment

L25–L29, L31–L32, L34, L35, L146, L149–L150, L178 (SWE unit), L198–L200, L200, L318–L319, L334, L344, L349, L352, L363–L364, L366, L395 (abbreviation), L462, L621, L623, L639, L663, L665, L670, L678, L766.

Part 4 — Summary of planned changes

- Forcing: remove the incorrect L180 statement; list per-model inputs in Table 2; report the completed common-input re-run of Experiment 1 in the Supplement (ranking among architectures and partition effect unchanged, so Stage 2 not triggered); state that the Experiment 1 comparison holds under common inputs and that comparisons in Experiments 2–5 are between deployed configurations; add limitations paragraph; foreground the within-architecture finding (G1).
- Terminology: define and apply PUB/PUR/PUC; replace "regional" with "hydroclimatic" generalisation throughout; delete regional-transfer claims (G4).
- Structure: move literature from Methods to Introduction; move background from Results to Methods; move computation time (Table 6) to an Appendix; remove redundancies (G3).
- Focus: integrate Experiment 5 into Fig. 3/Table 4 and Sect. 3; remove Sect. 4.2 from main text (corrected version in Supplement); move Fig. 4, example hydrographs and RGI categorisation out of the main text; move Tables S5–S6 to an Appendix (G5, G6).
- Statistics: drop the 0.5 NSE threshold and pBias bands; add paired per-basin difference analysis with effect sizes for partition contrasts and for Experiment 5; keep and explain the $\text{NSE} < 0$ threshold.
- Disclosures: Caravan version; streamflow-continuity criterion; hyperparameter table; temperature averaging in the PDD disaggregation; reasons for 10 trials in Experiment 5 and 36 trials for Graph WaveNet; definition of the high-mountain class and its exclusion of lowland glaciers.
- References: full audit; add all references suggested by the referee.
- Figures: redraw at publication quality; finer ticks and gridlines in Fig. 3.

References

- Acuña Espinoza, E., Loritz, R., Álvarez Chaves, M., Bäuerle, N., and Ehret, U.: To bucket or not to bucket? Analyzing the performance and interpretability of hybrid hydrological models with dynamic parameterization, *Hydrology and Earth System Sciences*, 28, 2705–2719, <https://doi.org/10.5194/hess-28-2705-2024>, 2024.
- 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.
- Álvarez Chaves, M., Acuña Espinoza, E., Ehret, U., and Guthke, A.: When physics gets in the way: an entropy-based evaluation of conceptual constraints in hybrid hydrological models, *Hydrology and Earth System Sciences*, 30, 629–658, <https://doi.org/10.5194/hess-30-629-2026>, 2026.
- Baste, S., Klotz, D., Acuña Espinoza, E., Bardossy, A., and Loritz, R.: Unveiling the limits of deep learning models in hydrological extrapolation tasks, *Hydrology and Earth System Sciences*, 29, 5871–5891, <https://doi.org/10.5194/hess-29-5871-2025>, 2025.

- Färber, C., Plessow, H., Mischel, S. A., Kratzert, F., Addor, N., Shalev, G., and Looser, U.: GRDC-Caravan: extending Caravan with data from the Global Runoff Data Centre, *Earth System Science Data*, 17, 4613–4625, <https://doi.org/10.5194/essd-17-4613-2025>, 2025.
- Feng, D., Beck, H., de Bruijn, J., Sahu, R. K., Satoh, Y., Wada, Y., Liu, J., Pan, M., Lawson, K., and Shen, C.: Deep dive into hydrologic simulations at global scale: harnessing the power of deep learning and physics-informed differentiable models (δ HBV-globe1.0-hydroDL), *Geoscientific Model Development*, 17, 7181–7198, <https://doi.org/10.5194/gmd-17-7181-2024>, 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.
- Gauch, M., Mai, J., and Lin, J.: The proper care and feeding of CAMELS: how limited training data affects streamflow prediction, *Environmental Modelling & Software*, 135, 104926, <https://doi.org/10.1016/j.envsoft.2020.104926>, 2021.
- Gleason, C. J. and Durand, M. T.: Remote sensing of river discharge: a review and a framing for the discipline, *Remote Sensing*, 12, 1107, <https://doi.org/10.3390/rs12071107>, 2020.
- 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.
- 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., Klotz, D., Hochreiter, S., and Nearing, G. S.: A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling, *Hydrology and Earth System Sciences*, 25, 2685–2703, <https://doi.org/10.5194/hess-25-2685-2021>, 2021.
- Liu, B., Yun, X., Pan, B., Xu, X., Gaffney, P. P. J., Lu, H., Luo, L., Sun, G., and Tang, Q.: Assess the impacts of climatic change and human activities on streamflow and floods by using a hybrid-physics-data (HPD) model: a case study in the Lancang-Mekong River Basin, *Journal of Hydrology: Regional Studies*, 61, 102763, <https://doi.org/10.1016/j.ejrh.2025.102763>, 2025a.
- Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T. Y., Weitzner, D., and Matias, Y.: Global prediction of extreme floods in ungauged watersheds, *Nature*, 627, 559–563, <https://doi.org/10.1038/s41586-024-07145-1>, 2024.
- 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.
- Song, Y., Sawadekar, K., Frame, J. M., Pan, M., Clark, M. P., Knoben, W. J. M., Wood, A. W., Lawson, K. E., Patel, T., and Shen, C.: Physics-informed, differentiable hydrologic models for capturing unseen extreme events, *Water Resources Research*, 62, e2025WR040414, <https://doi.org/10.1029/2025WR040414>, 2026.
- 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.