

Responses to the reviewers

Manuscript egusphere-2026-4749

A climate similarity-based transfer learning framework using global Caravan dataset for enhancing streamflow prediction in Ouémé River Basin (Benin, West Africa)

Response to Reviewer 2

Dear Reviewer, thank you for recognizing the potential value of transferring information to poorly gauged regions, and for pointing out the methodological weaknesses of the current manuscript. We accept that clearer reporting alone will not resolve all of these concerns. The revision will therefore add controlled comparisons and interpret the findings more cautiously.

Each comment is summarized below in italics and keeps the numbering of your report. Our replies follow in plain text.

Comment 3.1(a): Role of observed discharge

Comment summary: Clarify whether discharge is an input or a prediction target.

Response: We agree that the wording is ambiguous. In the intended rainfall–runoff setup, observed discharge is the prediction target, not a dynamic predictor. We will check this against the configurations that produced the reported results and list the meteorological inputs explicitly. We will also explain separately how Bonou discharge is used to calibrate the CI, since leaving discharge out of the LSTM inputs does not make donor selection independent of local observations. This calibration will not use the held-out target discharge.

Comment 3.1(b): Runoff and soil descriptors

Comment summary: Reconcile the description of runoff, groundwater and soil attributes with Table A1.

Response: Table A1 does include run_mm_syr (mean annual land-surface runoff) and dis_m3_pyr (mean annual natural discharge). We will make clear that these are the descriptors behind the reference to runoff conditions, and state where the data come from. We agree, however, that the reference to soil properties is wrong: groundwater-table depth is not a soil descriptor. We will correct the attribute categories and their descriptions.

Comment 3.1(c): Soil and land cover information

Comment summary: Justify the omission of soil and land-cover descriptors or evaluate their inclusion.

Response: We agree that climatic similarity does not capture every control on the rainfall–runoff transformation. We will document why the attributes were originally chosen and test a small set of soil and land-cover descriptors that are consistently available. The comparisons will separate their effect on donor selection from their effect as model inputs, instead of changing both at once. We will also check missing values, scaling and redundancy. Whether these descriptors improve transfer, or change the ranking of donor-set sizes, is something we will test rather than assume.

Comment 3.1(d): Streamflow conversion

Comment summary: Verify Equation (1), its implementation and the implications of any correction.

Response: We will check the drainage-area units and the discharge conversion in the preprocessing code itself, as described in our response to Reviewer 1, Comment 2. An incorrect unit in the text and an error in the input data are two different problems, and we will state which one applies. If the implementation is affected, we will regenerate the data concerned and repeat the affected analyses before reporting revised results.

Comment 3.1(e): Description of the target stations

Comment summary: Provide a comprehensive summary of the five stations and their data availability.

Response: We will add a table describing each station: drainage area, observation periods, number of valid daily observations, percentage of missing data in each split, mean precipitation, mean discharge, runoff ratio, and dominant soil and land-cover classes. We will state the periods and completeness criteria used for these calculations, and compute runoff ratios over matching precipitation and discharge periods. Before attributing differences between stations to short or event-poor records, we will quantify how much usable data, and how many flood events, each station actually has.

Comment 3.1(f): Attribute consistency and domain coverage

Comment summary: Harmonize donor and target attribute processing and use PCA or a similar analysis to assess their overlap.

Response: We agree that a shared framework does not guarantee identical processing. We will document dataset versions, basin boundaries, aggregation methods and unit conversions, and correct any inconsistencies between donor and target attributes. Attribute standardization and the PCA will be fitted on the donor data, and the target basins will then be projected into that same space. Closeness on the first two principal components cannot by itself establish hydrological equivalence or guarantee that transfer will succeed, so we will also compare the distributions of the main attributes.

Comment 3.2(a): Selection of the 200 donors

Comment summary: Explain whether the 200 catchments form the full eligible tropical subset or a further selection.

Response: We acknowledge that the selection is not documented well enough. We will report the Caravan version, the source of the Köppen–Geiger classification and the spatial rule used to assign basins to a class, as well as the data-availability requirements, quality filters and exclusions. A selection summary will give the number of catchments kept at each step and state clearly whether the 200 basins are the complete eligible subset. The final basin identifiers, their source datasets and their training, validation and test assignments will be released with reproducible code.

Comment 3.2(b): Rationale and alternative similarity indices

Comment summary: Justify climate-based selection within the tropical class, address aridity and compare alternative similarity definitions.

Response: We agree that the two-descriptor index needs a stronger justification and an empirical comparison. Basins in the same tropical class can still differ in precipitation amount and seasonality, so climate remains a reasonable selection criterion within that class. That alone does not show it is better than physiographic or combined criteria. Recent transfer-learning studies also stress catchment similarity and complementarity between donors (Ougahi and Rowan, 2026; Gao et al., 2026).

At a fixed donor-set size, we will compare the proposed CI with an equal-weight version, a climate version that includes aridity, and a combined climatic and physiographic index. We will report how the descriptors depend on one another, but a correlation alone will not be taken as proof that aridity adds no information. No data-dependent selection step will use test discharge. We will also discuss how the original CI depends on Bonou observations, and narrow our claims about ungauged basins accordingly.

Comment 3.2(c): Reproducibility of the index

Comment summary: Provide complete definitions of the normalized deviations, correlations, weights and final ranking.

Response: We will give the precipitation and seasonality formulas, the reference values and estimation periods, the normalization rules, and the exact pairs of observations used in the correlation analysis. We will distinguish correlations computed on pooled basin–year records from correlations computed within each donor basin, and explain which of these quantities set the weights. We will also state how negative or undefined correlations, missing years and temporal aggregation are handled.

We will check the calculation against the published equations and limit the discharge-based calibration to the target training period. Because the same annual Bonou discharge is reused for every donor, we will not treat the resulting correlations as independent hydrological evidence. The calibrated index will be compared with an equal-weight reference and judged on held-out predictive performance, not on its correlation with its own components. If these checks change the donor ranking, we will repeat the affected experiments.

Comment 3.3(a): Model specifications

Comment summary: Report sequence length, layer count, adaptation epochs and missing-target handling, and establish comparability of LSTM1–4.

Response: We will add a configuration table linked to the runs behind each result. It will list the number of recurrent layers, hidden size, sequence length, time steps contributing to the loss, dynamic and static inputs, dropout, batch size, epochs, stopping criteria and trainable parameters. Masking of missing targets and handling of missing inputs will be documented separately. We will check the target-station lists and resolve any differences between the manuscript and the released configurations. Where necessary, the comparisons meant to isolate the effect of pretraining will be rerun with harmonized data and evaluation protocols.

Comment 3.3(b): Hyperparameter search

Comment summary: Provide the search space and optimization procedure.

Response: We agree that the current description of manual tuning is not sufficient. We will report the values actually tried, separate fixed settings from tuned ones, and give the validation criterion, the checkpoint-selection rule and the computational budget. A limited manual search will not be described as exhaustive. In the revised comparisons, local and transferred models will have comparable tuning opportunities, and no tuning decision will use the test data.

Comment 3.3(c): Random initialization

Comment summary: Repeat experiments with multiple seeds and report performance variability.

Response: We agree that a single run cannot show whether modest performance differences are robust. The main model comparisons will be repeated with five training seeds, and we will report the resulting distributions and the paired differences at each station. For random donor selection, we will separate the variability across donor samples from the variability across initializations. When interpreting uncertainty, we will take into account that the target stations are not hydrologically independent. If we evaluate an ensemble, its performance will be reported separately from the distribution of individual-run scores.

Comment 3.3(d): Fine tuning strategy

Comment summary: Specify frozen and updated layers, learning-rate selection and the rationale for the adaptation strategy.

Response: For each adaptation experiment, we will state which modules are trainable and how many parameters they contain. With the head-only strategy currently described, the recurrent weights stay fixed, so adaptation changes the output mapping but not the internal temporal representation. We will explain how the learning rate was chosen on validation data and, as proposed in our response to Reviewer 1, Comment 3, compare head-only adaptation with unfreezing the recurrent layer. The donor-ranking results will be interpreted in light of the adaptation strategies we actually tested.

Comment 3.3(e): Conceptual hydrological benchmark

Comment summary: Include a conceptual rainfall–runoff model to assess the practical value of the framework.

Response: We agree that comparing LSTM variants alone does not show how the framework performs against a conventional hydrological model. We will add a GR4J benchmark for the five target stations, using consistent meteorological data, evaluation periods and admissible discharge observations. Potential evapotranspiration, model initialization, calibration criteria and parameter selection will be documented. The comparison will cover both overall performance and high-flow behavior. The previous application at Bonou alone cannot replace a benchmark run within the present experimental design.

Comment 3.4: Training records and evaluation duration

Comment summary: A long nominal training period does not demonstrate severe temporal scarcity, and the short test period limits flood-related conclusions.

Response: We agree that a 1996–2012 training interval is not, in itself, a short hydrological record. We will separate the problem of limited spatial coverage from that of temporal data scarcity, and report the observations actually usable at each station. Unsupported claims about short records will be removed. We will also add an experiment with shortened local training windows, to see how the benefit of transfer changes as less local information is available. The small number of target basins is a reason to evaluate transfer, but it does not show that local training is impossible, and the wider literature supports testing the benefits of multibasin learning (Kratzert et al., 2024).

We also accept that five test years give limited evidence about extreme events. We will check whether a longer evaluation period is available and add a complementary temporal evaluation, retraining the models for each partition and confining all selection to the corresponding training and validation data. The number and range of independent events covered will be reported. If the observations cannot support a longer usable evaluation, we will explain this limitation and restrict our conclusions to the events evaluated. In any case, no record length will be treated as sufficient on its own to validate performance for rare floods.

Comment 4.1: Simulation and forecasting

Comment summary: Distinguish retrospective simulation from operational forecasting and flood warning.

Response: We agree. The current experiments evaluate retrospective streamflow simulation driven by meteorological forcings. They do not establish operational forecasting skill at any specified lead time. We will revise the terminology in the abstract, methods, discussion, conclusions and short summary. Forecasting will be presented as a possible future application, one that would require dedicated experiments with defined issue times, horizons and real-time input availability. Claims that the current simulations already demonstrate operational flood-warning capability will be removed.

Comment 4.2: High flow and event metrics

Comment summary: Complement aggregate metrics with measures targeting high flows, peaks and flood events.

Response: We agree that aggregate scores alone do not show how well high flows are reproduced. We will add peak-flow bias, high-flow volume bias, errors on upper flow quantiles and peak-timing errors, along with event-based diagnostics. We will specify thresholds, event separation and matching rules, the treatment of incomplete records, and event counts; conclusions on timing will be limited by the daily resolution of the data. We will also distinguish the 95th non-exceedance percentile of daily flow from the high-flow volume bias metric FHV. In the revised interpretation, gains in overall fit and gains in flood-event behavior will be reported separately.