

## 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 1

Dear Dr Li, thank you for your constructive review, particularly your recommendations on separating donor-set size from similarity and ensuring consistency between the manuscript and the implementation.

For readability, the comments are summarized in italics, following the numbering of each reviewer's report. Our responses appear in plain text.

#### Comment 1: Donor set size and climatic similarity

*Comment summary: The cumulative donor sets confound size and similarity; fixed-size selection experiments and separate assessments of sampling and initialization variability are needed.*

**Response:** We agree. The present experiments compare three specific donor configurations but do not independently establish the effects of donor-set size and climatic similarity. We will compare the CI-selected set of 50 donors with repeated random selections of 50 donors from the same eligible pool, excluding the fixed validation and test basins. Target data, model inputs, training procedures and evaluation periods will be held consistent. Multiple donor samples and training seeds will allow us to distinguish donor-sampling variability from initialization variability and report paired performance differences at each station.

Because donor origin may also affect the comparison, we will examine source-dataset composition and include a source-matched comparison where the eligible pool permits it. We will describe the 50-donor set as the best configuration among those currently tested, rather than an established optimum. Even a positive fixed-size comparison would establish the value of CI selection at that size, not demonstrate that similarity universally outweighs volume. This distinction is consistent with evidence that diverse training data can also improve hydrological models (Fang et al., 2022). Performance-based selection and smaller donor sets would be supplementary analyses, not substitutes for the fixed-size comparison; any selection criterion would exclude the final test data.

#### Comment 2: Equations and streamflow units

*Comment summary: Correct the drainage-area units, the Spearman equation and the index notation, and determine whether the errors affect the calculations.*

**Response:** We acknowledge these inconsistencies. For discharge  $Q$  in  $\text{m}^3 \text{s}^{-1}$  and drainage area  $A$  in  $\text{km}^2$ , the conversion is  $q = 86.4 Q/A$  in  $\text{mm d}^{-1}$ . We will verify the units actually used in preprocessing, document a numerical conversion example and confirm consistency between Caravan and the target data. We will correct the squared rank-difference term in the Spearman equation and distinguish the number of paired observations from the number of climate descriptors in the weighting equations.

Eq. (2): the missing square is a typographical error only; the implementation uses `scipy.stats.spearmanr` (python packages SciPy), which also handles ties. We will check the implementation used for the reported results, including the treatment of tied ranks

#### Comment 3: Adaptation protocol and baseline comparability

*Comment summary: Clarify the parameters updated during fine-tuning, the organization of target data, preprocessing and checkpoint selection, and distinguish adaptation gains from pretraining gains.*

**Response:** We agree that the interpretation must match the scope of parameter updates. Under the protocol described in the manuscript, only the linear regression head is updated; the recurrent representations remain fixed. We will therefore describe this as adapting the mapping from pretrained features to local discharge, not relearning temporal representations. Restricting trainable parameters is a plausible way to limit local estimation demands, but its superiority cannot be assumed. We will include a targeted comparison between head-only adaptation and adaptation with the recurrent layer unfrozen, with learning rates selected using validation data.

We will reconcile the configurations associated with the reported results and document joint or station-specific training, the exact basin lists, sequence length, static-input integration, missing-data handling, input and target scaling, and checkpoint selection. The local baseline and transferred models will use comparable inputs, periods, loss evaluation and target-station organization. We will distinguish improvement over direct transfer, which measures local adaptation, from improvement over training from scratch, which assesses the contribution of pretraining. Discharge will be clearly identified as the prediction target rather than a meteorological forcing.

#### **Comment 4: Accuracy of references**

*Comment summary: Several cited studies do not support the corresponding claims, and the Frame et al. entry is duplicated.*

**Response:** We apologize for these errors. We will check the references against the specific claims they support, correct or remove the mismatched Song, Yao and Yang citations, and merge the duplicate Frame entries. Unsupported explanations, including the proposed groundwater explanation for station-specific behavior, will be removed or explicitly presented as hypotheses. We will also revise the discussion of extremes: the peak underestimation observed in our experiments should not be presented as a universal limitation of LSTMs. Frame et al. (2022, with the 2023 corrigendum) and Martel et al. (2025) provide evidence requiring a more qualified discussion.

#### **Comment 5: Duplicated sentence**

*Comment summary: Delete the duplicated sentence on page 8 at line 210.*

**Response:** We agree and will delete the duplicated sentence.

#### **Comment 6: Methods background**

*Comment summary: Shorten the general description of LSTMs and provide more study-specific implementation details.*

**Response:** We will condense the generic LSTM background and use the space to explain the model inputs, training stages, preprocessing and model-selection procedures specific to this study.

#### **Comment 7: Figure references**

*Comment summary: Change the references to Figure 5a and Figure 5b in Section 3.1 to Figure 4a and Figure 4b.*

**Response:** We will correct these references and check all figure and table citations throughout the manuscript.

#### **Comment 8: Model stages and Figure 6 sample sizes**

*Comment summary: Distinguish pretrained and fine-tuned states and clarify the training, validation and test samples represented in Figure 6.*

**Response:** We will distinguish pretrained configurations, direct-transfer predictions and locally adapted models consistently in the text and figures. After checking the basin lists against the associated runs, the Figure 6 caption will identify the 50, 100 and 150 training-donor configurations and the fixed validation and test sets of 30 and 20 basins. We will reconcile the selection script, released lists and figure labels so that each displayed distribution has an unambiguous sample definition.