https://doi.org/10.5194/egusphere-2026-2950
Preprint. Discussion started: 16 June 2026

Technical note: Regional fine-tuning of LSTMs for improved streamflow predictions in ungauged catchments

Ashkan Shokri[1], James C. Bennett[1], David E. Robertson[1]

[1]Commonwealth Scientific and Industrial Research Organisation (CSIRO), Clayton 3168, Australia

Correspondence to: Ashkan Shokri (ash.shokri@csiro.au)

Abstract. Predicting streamflow in ungauged basins (PUB) remains a central challenge in hydrology. Long short-term memory (LSTM) networks trained on large samples of catchments ("global" LSTMs) have emerged as a state-of-the-art approach for PUB, outperforming conceptual rainfall–runoff models with traditional regionalisation approaches. However, global LSTMs are spatially agnostic, relying solely on static catchment attributes to differentiate regional hydrological behaviour. This study introduces Regionalised Fine-Tuning (ReFT), a strategy that adapts a pretrained global LSTM to the region surrounding each ungauged target catchment by fine-tuning on a spatially weighted set of donor catchments using an inverse-distance weighting scheme. ReFT is evaluated on 218 catchments from the CAMELS-AUS dataset under a spatial out-of-sample cross-validation framework, comparing two fine-tuning configurations: updating all model parameters versus updating only the prediction head while keeping the recurrent backbone frozen. ReFT improves Nash–Sutcliffe Efficiency relative to the base global LSTM in more than 66% of catchments, with the largest gains occurring for catchments of moderate baseline performance. The ReFT framework combines the broad process generalisation of large-sample deep learning with the local specificity of regional adaptation, providing an efficient route to improved streamflow predictions in data-sparse regions.

1 Introduction

Prediction in Ungauged Basins (PUB) remains one of the most significant challenges in hydrology (Razavi and Coulibaly, 2013). Historically, PUB has been addressed through donor-based regionalization approaches, particularly for Conceptual Rainfall–Runoff (CRR) models. In these approaches, model parameters for an ungauged catchment are inferred from one or more hydrologically similar gauged donor basins, selected based on geographic proximity, climatic similarity, or physiographic attributes (Demirel et al., 2024; Gebeyehu et al., 2023; Hrachowitz et al., 2013; Huang and Liang, 2006; Prakash et al., 2025; Razavi and Coulibaly, 2013; Spence et al., 2013; Wagener and Montanari, 2011; Yin et al., 2025). Perhaps the most widely applied of these regionalisation methods is the simple 'nearest neighbour' approach, where predictions in an ungauged basin are generated by a CRR with parameters transferred from the closest gauged basin. By restricting information transfer to a subset of comparable catchments, donor-based methods attempt to exploit spatial coherence and regional similarity, allowing shared but often unobserved hydrological controls, such as geology or drainage organization, to be implicitly represented.

https://doi.org/10.5194/egusphere-2026-2950
Preprint. Discussion started: 16 June 2026

Recent advances in Deep-Learning (DL) methods for hydrology, particularly long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), have demonstrated promising performance in hydrological modelling for ungauged basins (Kratzert et al., 2019; Lee et al., 2025; Heudorfer et al., 2026; Shokri et al., 2026). When trained on hundreds (sometimes thousands) of catchments, using a combination of static catchment attributes (e.g., mean annual precipitation, soil characteristics, or land cover) and dynamic predictors (e.g. precipitation time series), LSTMs are able to predict streamflow in ungauged basins more efficiently than conceptual rainfall-runoff (CRR) models with commonly used regionalisation methods. We will refer to an LSTM that is tuned to hundreds of gauges - e.g. from across a continent - as a 'global' LSTM to distinguish this type of model from a 'finetuned' LSTM, which takes additional information from localised gauges, as described below. Despite their success, global LSTMs are inherently spatially agnostic. During training, the training dataset is treated as a homogeneous pool of information, where the model's capacity for regional differentiation relies exclusively on a fixed set of static catchment attributes.

For PUB with CRRs, Shokri et al., (2026) attempted to combine the benefits of localised information with information from a wider array of catchments by calibrating the GR4J CRR using a multi-site pooling approach. This method pooled hundreds of GR4J models at individual catchments and weighted the calibration objective based on inverse distance from each ungauged catchment. This improved the performance of GR4J for PUB over a nearest-neighbour regionalisation approach but still did not outperform a global LSTM.

For predictions in gauged catchments, global LSTMs are often 'finetuned' to each gauge, which usually substantially improves the predictions of the global model (Shokri et al., 2026). Finetuning refines a global LSTM's internal parameters to the unique behaviours of an individual catchment, resulting in highly accurate long-term historical simulations. The success of finetuning shows that highly localised information can improve the performance of global LSTMs.

In this study, we combine these three threads:

1. A global LSTM tuned to hundreds of catchments
2. Regionalised Fine-Tuning (ReFT) of the global LSTM from (1) to a region (rather than to individual gauges), enabling PUB
3. Using a form of inverse-distance weighting to emphasise local information for ungauged basins in the ReFT in (2)

The ReFT strategy attempts to inject implicit latent spatial context into the pretrained global LSTM. Unlike feature-based approaches that rely solely on static catchment descriptors, ReFT explicitly constrains the transfer of hydrological information across space.

The primary objective of this study is to evaluate whether ReFT improves predictions in ungauged basins over a global LSTM. The method is tested on the CAMELS-AUS dataset to determine whether distance-weighted information transfer from gauged neighbours can substitute for local streamflow observations, improving predictive performance under a Spatial out-of-Sample (SooS) evaluation framework.

https://doi.org/10.5194/egusphere-2026-2950
Preprint. Discussion started: 16 June 2026

2 Methods

2.1 Data and Catchment Attributes

The CAMELS-AUS dataset (Fowler et al., 2021) provides daily hydrometeorological time series and catchment attributes for 222 minimally impacted catchments across Australia. We removed 4 catchments due to short data records, leaving 218 catchments. We employ two categories of inputs to the LSTMs:

Dynamic Predictors: Daily meteorological forcings, specifically precipitation from the Australian Gridded Climate Data (AGCD) and potential evapotranspiration (PET) from the SILO database.

Static Attributes: Landscape characteristics used to differentiate catchment behaviour, including climatic indices and geomorphological features (Table 1).

Table 1 Static and quasi static features for LSTM Model

Category	Predictor	Description
Climatic and Precipitation Characteristics (static)	p_mean	Mean Annual Precipitation
	pet_mean	Mean Annual Potential Evapotranspiration
	Aridity	Aridity (Mean Annual PET/Mean Annual Precipitation)
	p_seasonality	Precipitation Seasonality
	high_prec_freq	Frequency of High-Precipitation Days (≥ 5 times mean annual)
	high_prec_dur	Average Duration of High Precipitation Events
Catchment and Geomorphological Characteristics (quasi-static)	catchment_area	Catchment Area
	mean_slope_pct	Catchment Mean Slope
	prop_forested	Proportion of Catchment Occupied by Forest
	Upsdist	Maximum Flow Path Length Upstream
	Strdensity	Ratio of Total Length of Streams to Catchment Area
	Strahler	Strahler Stream Order at Gauging Station

2.2 The base model: global LSTM

A global LSTM serves as the foundation of the approach, trained on all available catchments simultaneously to learn a generalized rainfall–runoff representation. The architecture follows Kratzert et al. (2018), comprising an input layer, a recurrent LSTM layer with a hidden state size of 256, and a dense output layer.

The model receives sequences of 365 days of dynamic meteorological forcings (i.e. precipitation and PET), x_t, and static catchment attributes a as forcing data to predict daily streamflow $\hat{y}_t$. Model parameters θ_{global} are optimized by minimizing the smooth-joint Nash-Sutcliffe Efficiency (NSE) loss (Kratzert et al., 2019) over the training set.

https://doi.org/10.5194/egusphere-2026-2950
Preprint. Discussion started: 16 June 2026

2.3 Regionalised Fine-Tuning (ReFT)

For an ungauged target catchment T, we denote a set of N donor catchments as $\mathcal{D}_T = \{d_1, d_2, \ldots, d_n, \ldots, d_N\}$. To account for the relevance of each donor catchment to the target basin, an inverse distance weighting (IDW) scheme is employed, so that the contribution of each donor catchment to the fine-tuning objective decreases with increasing spatial separation from the target catchment. The weight assigned to donor catchment n is defined as:

$$w_n = \frac{d_{n,T}^{-\alpha}}{\sum_{j=1}^{N} d_{j,T}^{-\alpha}} \quad (1)$$

where $d_{n,T}$ is the distance between the centroids of target catchment T and donor catchment n, and $\alpha \geq 1$ is a parameter controlling how rapidly donor influence decreases with increasing distance. Larger values of α place greater emphasis on nearby catchments, whereas smaller values allow information from distant catchments to contribute more to the fine-tuning process. We trialled values of α = 1, α = 2 and α = 3 (not shown). These performed similarly, and we choose α = 2 for this study.

The global model parameter set θ_{global} is updated to obtain a set of target-specific parameters, θ_T, by minimizing a spatially weighted loss function evaluated across all donor catchments. Specifically, the ReFT loss function is defined as:

$$\mathcal{L}_{\text{ReFT}}(\theta) = \sum_{n=1}^{N} w_n \cdot \mathcal{L}_{\text{NSE}}(y_n, \hat{y}_n(\theta)) \quad (2)$$

where $\mathcal{L}_{\text{NSE}}$ is the Nash–Sutcliffe Efficiency loss, y_n denotes the observed streamflow for donor catchment n, and $\hat{y}_n(\theta)$ represents the corresponding LSTM predictions.

By minimizing this objective function, the LSTM adapts its internal states and parameter values to emphasize hydrological dynamics that are characteristic of the region surrounding the target catchment. In effect, ReFT transfers regional hydrological knowledge from neighbouring gauged basins into the pretrained model, producing a target-specific LSTM that is better suited to the local hydrological regime despite the absence of direct streamflow observations for the target catchment.

ReFT is computationally demanding, and scales with the value of N. To make the process more tractable, we experimented with using only the nearest 5%, 10% and 20% of all available donor catchments. We found performance did not materially change (not shown) – very likely because the weights applied to increasingly distant catchments are very small. To minimise computational cost in this study, for each target catchment T we restrict N to the nearest 5% of available donor catchments. Under spatial cross-validation (see Experimental Design section, below), we have ~164 donor catchments available, meaning we use the nearest $164 \times 5\% \approx 9$ catchments to inform the ReFT.

2.4 Fine-Tuning Configurations

Two parameter-update strategies are evaluated within the ReFT framework:

1) Full-parameter fine-tuning, in which all model parameters are updated during finetuning.

https://doi.org/10.5194/egusphere-2026-2950
Preprint. Discussion started: 16 June 2026

2) Prediction-head fine-tuning, in which the recurrent LSTM layers are frozen and only the final fully connected prediction layer is updated. In this setting, the global model retains its temporal feature representations, while regional adaptation is restricted to the mapping between latent states and streamflow predictions.

2.5 Experimental Design

Model performance in ungauged catchments is evaluated using Spatial out-of-Sample (SooS) cross-validation, following the partitioning strategy established by Shokri et al. (2026). The dataset of 218 catchments was divided into four spatially distinct folds. The splits are randomised without replacement, with the exception that nested catchments are blocked together to ensure they were always in the same fold. This prevents hydrologically dependent catchments from being split between training and validation sets, thereby avoiding potential data leakage. Each fold is treated as a validation set in turn, with the remaining three folds forming the training dataset, ensuring that every catchment is evaluated under pseudo-ungauged conditions.

For each validation fold, donor catchments are drawn exclusively from the training folds to maintain independence between training and validation data. Both the global LSTM model and ReFT are trained only on the training folds. After fine-tuning, predictions are generated for the validation catchments using the target-specific model parameters and the performance metrics are calculated.

To benchmark the ReFT-LSTM model it is compared against three reference approaches:

- The global LSTM (Base) described in Sect. 2.2.
- A regionalised GR4J CRR, for which parameters for ungauged catchments are estimated by a multi-site objective from donor catchments with the same IDW scheme as Eq. (1) where $\alpha = 2$ (see Shokri et al. (2026) for detailed description).
- The AWRA-L land surface model (Frost and Shokri, 2021; details in Appendix A).

To ensure robustness against stochastic initialization, all LSTM experiments are repeated across 10 random seeds, and median performance metrics are reported.

3 Results

Figure 1 shows that all the LSTM-based approaches trialled in this study substantially outperform AWRA-L and the regionalised GR4J in all but the most poorly performing catchments. This is consistent with results reported by Shokri et al. (2026) for Australia and for a number of studies elsewhere (e.g. Kratzert et al., 2019; Heudorfer et al., 2026).

ReFT shows consistent improvement in PUB performance over the global LSTM, with performance increasing in most catchments (Fig. 1). ReFT improves performance over the base LSTM in more than 66% of catchments, with gains visible across much of the distribution. Beyond approximately the 90% exceedance probability, however, the ReFT curves drops slightly below that of the base LSTM, indicating a minor reduction in performance for the lowest-performing catchments. In

https://doi.org/10.5194/egusphere-2026-2950
Preprint. Discussion started: 16 June 2026

this same range, the most difficult-to-model catchments, the regionalised GR4J curve remains above all LSTM-based approaches, suggesting greater robustness of the conceptual model in basins with weak hydrological predictability.

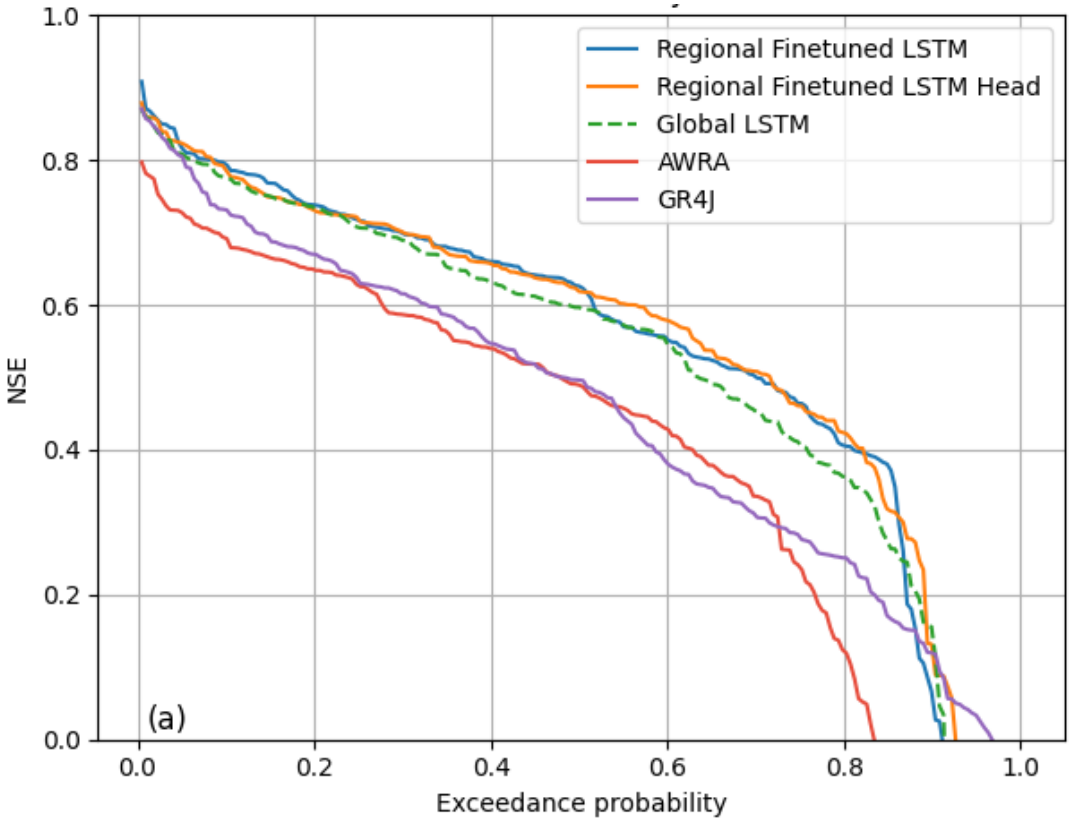

Figure 1 Exceedance probability plots of Nash–Sutcliffe Efficiency (NSE) for ungauged catchments under Spatial out-of-Sample (SooS) conditions, comparing the base continental LSTM with the corresponding Regionalised Fine-Tuned (ReFT) LSTM.

The largest performance gains occur near the centre of the exceedance distribution, indicating that ReFT is most effective in catchments with moderate baseline performance.

Figure 2 and Figure 3 Spatial comparison of model performance in ungauged catchments under SooS cross-validation, showing catchments where ReFT with only predicting head finetuning achieve higher NSE relative to the base LSTM. illustrates the spatial distribution of ReFT impacts on LSTM performance. Overall, prediction-head-only fine-tuning improved performance across a larger number of catchments than full-parameter fine-tuning. In contrast, updating all model parameters produced a more polarized response: more catchments experienced substantial improvements (NSE increase > 0.1), but a larger number also exhibited substantial performance degradation (NSE decrease > 0.1).

https://doi.org/10.5194/egusphere-2026-2950
Preprint. Discussion started: 16 June 2026

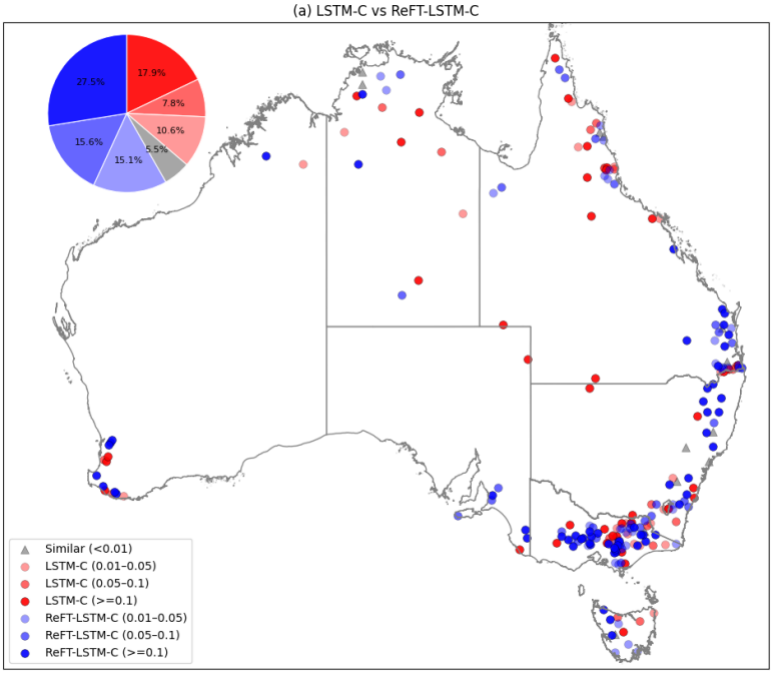

Figure 2 Spatial comparison of model performance in ungauged catchments under SooS cross-validation, showing catchments where ReFT with full parameter finetuning achieve higher NSE relative to the base LSTM.

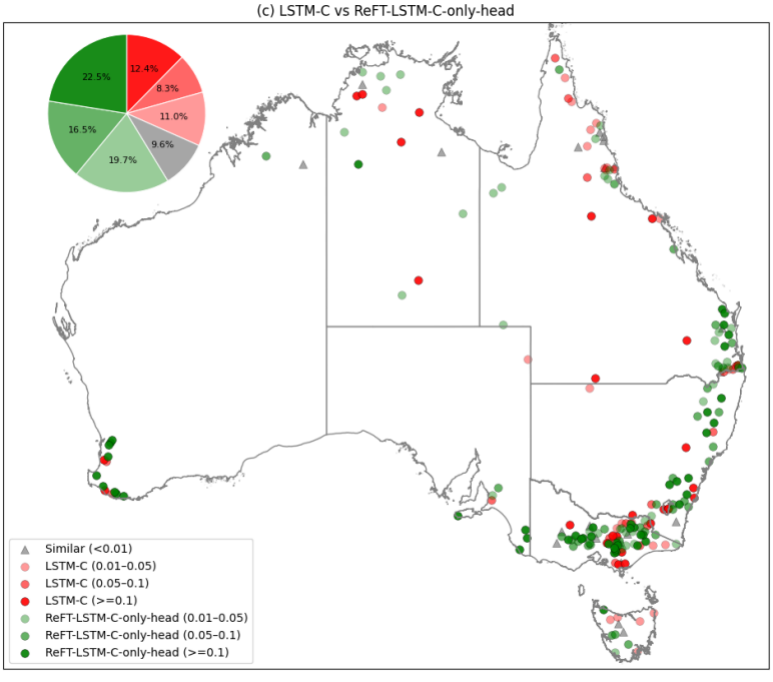

Figure 3 Spatial comparison of model performance in ungauged catchments under SooS cross-validation, showing catchments where ReFT with only predicting head finetuning achieve higher NSE relative to the base LSTM.

https://doi.org/10.5194/egusphere-2026-2950
Preprint. Discussion started: 16 June 2026

Prediction-head-only fine-tuning produced more moderate performance changes and fewer cases of large NSE declines, indicating more consistent behaviour across catchments. In contrast, full-parameter fine-tuning exhibited greater variability, with larger potential gains accompanied by increased performance deterioration in some regions. We believe that more consistent benefits are likely to be favoured for most PUB applications, meaning that head-only fine-tuning is overall preferable to full-parameter fine-tuning.

Figure 4 compares the behaviour of full-parameter and prediction-head-only fine-tuning strategies catchment-by-catchment. Among the catchments where the two strategies disagree by more than ±0.05 in NSE, head-only fine-tuning is the better performer in 26.1% of catchments compared with 20.2% for full-parameter fine-tuning, and at a threshold of ±0.10 the gap widens to 17.0% versus 9.2%. At very small thresholds (±0.01), full-parameter fine-tuning shows a marginal edge (48.2% versus 39.4%), but these differences are unlikely to be practically meaningful. The largest disagreements (red points with |ΔNSE| > 0.10) are concentrated among catchments with low NSE values, indicating that the two strategies diverge most in the more challenging catchments where the global LSTM already struggles and, in these cases, head-only fine-tuning is the more reliable choice.

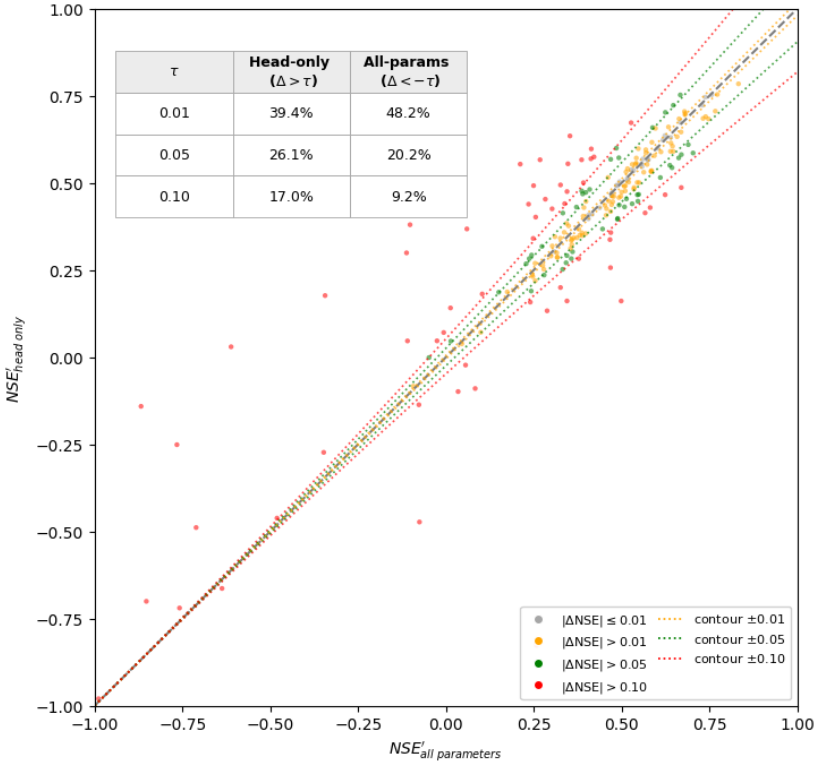

Figure 4 Improvement in transformed NSE (NSE') over the base LSTM model for ReFT full-parameter and prediction-head-only fine-tuning, where $NSE' = \frac{NSE}{2-NSE}$. This transformation is used to improve visualization in the scatter plot. Points are coloured by the absolute difference in NSE between the two strategies (|ΔNSE|), and dashed contour lines indicate ±0.01, ±0.05 and ±0.10 offsets from the 1:1 line. The inset table reports, for each threshold τ, the percentage of catchments where head-only fine-tuning exceeds full-parameter fine-tuning by more than τ (Δ > τ) and the percentage where it falls below by more than τ (Δ < −τ).

https://doi.org/10.5194/egusphere-2026-2950
Preprint. Discussion started: 16 June 2026

The relationship between the two approaches is further illustrated by the scatter plot of NSE improvements relative to the base LSTM. A positive association is observed between the two strategies, indicating that catchments benefiting from full-parameter ReFT generally also benefit from prediction-head-only updates.

4 Discussion

We have demonstrated that ReFT improves predictive performance in most ungauged catchments relative to the base global LSTM. A key strength of the ReFT approach lies in its ability to preserve the benefits of large-sample learning and build on this with relevant regional information. ReFT contrasts with traditional regionalization techniques applied to CRRs such as GR4J, where model structures are fixed and parameters must be inferred entirely from donor information. The ReFT framework therefore integrates the broad process generalization of a global model with the local specificity of regional adaptation, effectively combining the advantages of both approaches.

The two fine-tuning configurations evaluated within ReFT exhibited distinct adaptation characteristics. Full-parameter fine-tuning produced larger improvements in some catchments but also greater performance degradation in others, whereas prediction-head-only fine-tuning yielded more moderate and consistent changes. This pattern is consistent with the interpretation that restricting updates to the prediction head preserves the temporal feature representations learned by the global LSTM, while still allowing the mapping from latent states to streamflow to be adapted regionally. Full-parameter fine-tuning, in contrast, permits broader adaptation but is more susceptible to overfitting on the relatively small pool of weighted donor catchments. The robustness advantage of head-only fine-tuning is most apparent in the lower tail of the performance distribution, suggesting it is the safer default in catchments where the global model already struggles.

While more complex similarity measures could be used, spatial distance captures large-scale gradients in climate and landscape properties that are known to strongly influence hydrological response. Physical and climatic similarity between catchments is also already represented within the LSTM through the static attributes provided as inputs (Table 1); spatial distance therefore contributes a distinct form of information, capturing spatial coherence in unobserved or imperfectly represented controls such as geology, drainage organisation, or sub-grid climatic structure. Distance is also simple to measure and implement, requires no prior decisions about which attributes to weight or how to scale them, and remains well-defined in data-sparse settings where attribute information may be incomplete or unreliable. Defining donor weights from a combination of attributes could plausibly improve prediction further, but identifying which attributes are informative, and how they should be weighted, would require a dedicated analysis beyond the scope of this technical note.

The improvements exhibited by ReFT may in part be due to a relatively small number of catchments in the CAMELS-AUS dataset, and the spatial disparity between gauges, both of which may limit the performance of the global LSTM in ungauged basins. Australia's streamflow gauges are overwhelmingly situated along the east coast, with a few clusters on the southwest and north coasts. The interior is very sparsely gauged. This contrasts with other continental datasets used to develop LSTMs in the literature (e.g. CAMELS-US), which often have hundreds more gauges that are much more uniform in their spatial

https://doi.org/10.5194/egusphere-2026-2950
Preprint. Discussion started: 16 June 2026

distribution. The effectiveness of ReFT may reduce compared to global LSTMs when trained on densely gauged regions such as the USA or Europe. Future testing on different datasets will inform the efficacy of ReFT in other settings.

Despite the improvements exhibited by ReFT, several challenges remain. While ReFT substantially enhances median and upper-quantile performance, limitations persist in the lower tail of the performance distribution, corresponding to catchments with poor predictability. In these difficult basins the regionalised GR4J model occasionally outperforms the LSTM-based approaches including the ReFT-LSTMs. This behaviour suggests that, in environments characterized by weak hydrological signals or strongly non-linear threshold responses, the structural constraints embedded in conceptual models may provide a stabilizing effect that limits unphysical extrapolation or overfitting. This complementarity also suggests that hybrid or ensemble strategies combining ReFT-LSTM with regionalised CRRs in difficult basins may be worth exploring.

A practical consideration is computational cost. ReFT requires a separate fine-tuning step for each target catchment, which scales linearly with the number of ungauged sites of interest. The use of a restricted donor pool (the nearest 5% of available catchments in this study) and the option of head-only fine-tuning both substantially reduce this cost without materially affecting performance, but operational deployment at national scale would still require careful workflow design.

5 Conclusion

This paper introduces Regional Fine Tuning (ReFT), a strategy for LSTMs that bridges the gap between large-sample deep learning and more localised information for predictions in ungauged basins. ReFT works by implementing an inverse-distance weighting scheme within the LSTM's fine-tuning phase. ReFT transforms spatially agnostic global models into spatially aware regional specialists, specifically:

- ReFT successfully transfers latent hydrological information from gauged donors to ungauged targets, resulting in similar or improved NSE scores for >66% of study catchments relative to the base continental LSTMs, with the largest gains observed in catchments with moderate baseline performance.
- Applying ReFT only to the prediction-head of the global LSTM provides slightly more robust performance than update all parameters in the LSTM, particularly in lower-performing catchments.

The framework offers an efficient mechanism for improving streamflow predictions in data-sparse regions, supporting more reliable applications in water resource assessment and historical reconstruction.

Code and data availability

The datasets used in this study are publicly available as follows: The AWRA-L dataset is available from the Australian Bureau of Meteorology (BoM). For access, visit the Australian Water Outlook website: https://awo.bom.gov.au/. SILO provides long-term climate datasets, including rainfall and potential evapotranspiration, from the Queensland Government. Access the SILO database at: https://www.longpaddock.qld.gov.au/silo/. The CAMELS-AUS dataset, including hydrometeorological timeseries

https://doi.org/10.5194/egusphere-2026-2950
Preprint. Discussion started: 16 June 2026

and catchment attributes, is available through Earth System Science Data. The dataset can be accessed via: https://doi.org/10.5194/essd-13-3847-2021. The code developed for this study is available upon request from the corresponding author due to licencing requirements.

Author contributions

AS, JCB, and DER conceived and designed the study. AS developed the models, performed the analyses, and wrote the manuscript. JCB and DER contributed to the interpretation and discussion of the results and assisted with reviewing and proofreading the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

This research was conducted on the traditional lands of the Boonwurrung people of the Kulin Nation. We pay our respects to their Elders past and present. We also acknowledge the Traditional Owners of the catchments used in this study. We acknowledge the use of artificial intelligence tools to assist with language refinement and proofreading of this manuscript. All AI-generated suggestions were reviewed and edited by the authors, who take full responsibility for the content of the manuscript.

References

Demirel, M. C., Koch, J., Rakovec, O., Kumar, R., Mai, J., Müller, S., Thober, S., Samaniego, L., & Stisen, S. (2024). Tradeoffs Between Temporal and Spatial Pattern Calibration and Their Impacts on Robustness and Transferability of Hydrologic Model Parameters to Ungauged Basins. *Water Resources Research*, *60*(1). https://doi.org/10.1029/2022WR034193

Fowler, K. J. A., Acharya, S. C., Addor, N., Chou, C., & Peel, M. C. (2021). CAMELS-AUS: Hydrometeorological time series and landscape attributes for 222 catchments in Australia. *Earth System Science Data*, *13*(8). https://doi.org/10.5194/essd-13-3847-2021

Gebeyehu, B. M., Jabir, A. k., Tegegne, G., & Melesse, A. M. (2023). Reliability-weighted approach for streamflow prediction at ungauged catchments. *Journal of Hydrology*, *624*. https://doi.org/10.1016/j.jhydrol.2023.129935

Heudorfer, B., Gupta, H., Dolich, A., & Loritz, R. (2026). Better data or better architecture? Improving deep-learning-based prediction in ungauged basins. *EGUsphere*, *2026*, 1–30. https://doi.org/10.5194/egusphere-2026-1965

https://doi.org/10.5194/egusphere-2026-2950
Preprint. Discussion started: 16 June 2026

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V., Hughes, D. A., Hut, R. W., Montanari, A., Pande, S., Tetzlaff, D., … Cudennec, C. (2013). A decade of Predictions in Ungauged Basins (PUB)-a review. In *Hydrological Sciences Journal* (Vol. 58, Number 6). https://doi.org/10.1080/02626667.2013.803183

Huang, M., & Liang, X. (2006). On the assessment of the impact of reducing parameters and identification of parameter uncertainties for a hydrologic model with applications to ungauged basins. *Journal of Hydrology*, *320*(1–2). https://doi.org/10.1016/j.jhydrol.2005.07.010

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11). https://doi.org/10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12). https://doi.org/10.5194/hess-23-5089-2019

Lee, J., Chung, E. S., Kim, S., & Kim, D. (2025). Streamflow forecasting in ungauged basins with CNN-LSTM and radar-based precipitation. *Journal of Hydro-Environment Research*, *60–61*. https://doi.org/10.1016/j.jher.2025.100666

Prakash, H., Pandey, K. K., & Soni, P. (2025). Peak discharge estimation for ungauged basins: a review. *Journal of Water and Climate Change*, *16*(11), 3483–3507. https://doi.org/10.2166/wcc.2025.153

Razavi, T., & Coulibaly, P. (2013). Streamflow Prediction in Ungauged Basins: Review of Regionalization Methods. *Journal of Hydrologic Engineering*, *18*(8). https://doi.org/10.1061/(asce)he.1943-5584.0000690

Shokri, A., Bennett, J. C., Robertson, D. E., Perraud, J.-M., Frost, A. J., & Lehmann, E. A. (2026). Better continental-scale streamflow predictions for Australia: LSTM as a land surface model post-processor and standalone hydrological model. *Hydrology and Earth System Sciences*, *30*(3), 757–777. https://doi.org/10.5194/hess-30-757-2026

Spence, C., Whitfield, P. H., Pomeroy, J. W., Pietroniro, A., Burn, D. H., Peters, D. L., & St-Hilaire, A. (2013). A review of the Prediction in Ungauged Basins (PUB) decade in Canada. In *Canadian Water Resources Journal* (Vol. 38, Number 4). https://doi.org/10.1080/07011784.2013.843867

Wagener, T., & Montanari, A. (2011). Convergence of approaches toward reducing uncertainty in predictions in ungauged basins. *Water Resources Research*, *47*(6). https://doi.org/10.1029/2010WR009469

Yin, H., Zhao, L., Zhu, M., & Zhang, Y. (2025). Runoff prediction in gauged and ungauged basins using Transformer-XAJ model. *Journal of Hydrology*, *662*. https://doi.org/10.1016/j.jhydrol.2025.133954

https://doi.org/10.5194/egusphere-2026-2950
Preprint. Discussion started: 16 June 2026

295

Appendix A: AWRA-L streamflow predictions

The Australian Water Resource Assessment Landscape (AWRA-L) is a land surface model (LSM) that produces simulations on a 0.05° grid across Australia, with national historical outputs available from 1911 onward. AWRA-L v7 is somewhat unusual among LSMs in that it is partially calibrated. Calibration is performed using data from 295 catchments over the period 1981–2011, employing an objective function that combines multiple observational constraints: GRACE terrestrial water storage (50%), streamflow (35%), satellite-based soil moisture (7.5%), evapotranspiration (2.5%), and vegetation fraction (5%), aggregated across all catchments (Frost and Shokri, 2021). The model has been extensively validated against a wide range of observational datasets (Frost et al, 2021). AWRA-L simulates multiple hydrological variables; in this study, only total runoff (Q_{tot}) was used. In AWRA-L, Q_{tot} is derived from surface flow, baseflow, and interflow, which are routed at the grid-cell scale through a conceptual surface water store to represent delayed stormflow responses. However, the model does not include channel routing, as grid cells are independent and lack lateral flow, which can lead to inaccuracies in large catchments with travel times exceeding one day. Streamflow at catchment outlets was therefore estimated by area-weighted aggregation of Q_{tot} across grid cells within each catchment, without accounting for in-stream routing, losses, overbank flow, or channel storage. The resulting catchment-scale runoff time series was used as a benchmark in the current study.

300

305

310