Catchment Filtering is not a Neutral Step: Missing-Data Thresholds Alter Basin Composition and Regional LSTM Performance in CAMELS-BR

Thiago M. Osawa [1] Dário H. Hossoda [2], Elton V. Escobar-Silva [3]& Jose Rodolfo S. Martins [4]*

Keywords – Hydrological data quality; basin representativeness; deep learning.

INTRODUCTION

Artificial intelligence (AI) and big data have become increasingly important in disaster management, especially for flood forecasting, early warning systems, and hydrological risk assessment. Regional deep learning (DL) models, particularly Long Short-Term Memory (LSTM) networks, can learn rainfall–runoff relationships from large catchment samples and support streamflow prediction in data-scarce or ungauged basins (Kratzert *et al.*, 2019; Tripathy and Mishra, 2024). However, their reliability depends not only on model architecture, but also on the quality, completeness, and representativeness of the hydrological data used for training (Snieder and Khan, 2025; Heudorfer and Loritz, 2026).

In Brazil, Catchment Attributes and Meteorology for Large-sample Studies in Brazil (CAMELS-BR) is a key dataset for regional hydrological modeling. It provides daily streamflow records, basin-mean meteorological forcings, and catchment attributes for 897 basins (Chagas *et al.*, 2020). The release of CAMELS-BR v1.2 expanded the available hydrometeorological time series, allowing regional models to use more recent data. Nevertheless, streamflow availability is uneven over time. Initial analyses indicate that the historical period from 1989 to 2009 provides the highest spatial coverage, with 882 of the 897 basins retained under a 5% missing-data threshold. In contrast, extending the analysis period to 2024 reduces the number of usable basins, especially under stricter missing-data criteria.

This trade-off is particularly important for regional DL models, whose performance depends on both the length of the training records and the diversity of catchments used for learning (Snieder and Khan, 2025). Longer and more recent records may capture additional hydroclimatic variability, but they may also reduce the number of eligible basins when data completeness is uneven over time. Conversely, shorter periods with higher data availability may preserve a larger and more representative catchment sample. Moreover, relaxing the missing-data threshold may not only increase the number of basins but also modify the spatial and hydroclimatic composition of the dataset.

Therefore, this study evaluates the sensitivity of regional streamflow modeling in CAMELS-BR to two methodological decisions: the training period and the maximum allowed percentage of missing streamflow data. Specifically, we investigate whether extending the training period to 2024 compensates for the reduction in catchment availability and whether relaxing the missing-data threshold improves model performance by increasing the number and diversity of basins. By addressing these questions, the study contributes to more transparent and reliable AI-driven hydrological modeling for disaster risk management and flood-related decision support.

METHODOLOGY

A regional rainfall–runoff model was implemented using the NeuralHydrology framework (Kratzert *et al.*, 2022). For each experimental setup, a single regional LSTM model was trained using the

1*) Department of Hydraulic and Environmental engineering of University of São Paulo, thiago.osawa@usp.br
2) Department of Hydraulic and Environmental engineering of University of São Paulo, dario.hossoda@usp.br
3) National Center for Monitoring and Early Warning of Natural Disasters (Cemaden), elton.escobar@cemaden.gov.br
4) Department of Hydraulic and Environmental engineering of University of São Paulo, scarati@usp.br

Compute Unified Device Architecture (CUDA)-accelerated LSTM implementation. The model comprised one LSTM layer with 256 hidden units, an output dropout rate of 0.4, and an input sequence length of 270 days. It predicted daily streamflow at the final step of each input sequence. CAMELS-BR meteorological forcings were used as dynamic inputs, whereas the static catchment attributes available in CAMELS-BR were used as catchment descriptors.

The model was trained with the Adam optimizer and a Nash–Sutcliffe efficiency (NSE)-based loss function. The initial learning rate was set to 5×10^{-4}, with a dynamic scheduler reducing it by a factor of 0.5 after three epochs without validation improvement. Training was limited to 30 epochs, with early stopping after a minimum of eight epochs and a patience of six epochs. The batch size was 128. Validation was performed after each epoch, and the final test evaluation used the checkpoint with the highest median validation NSE.

Four experiments were designed to isolate the effects of the missing-data threshold and training-period length (Table 1). Experiments E1, E2, and E3 used the same period, from 1990-09-01 to 2024-08-31, but applied missing-data thresholds of 5%, 10%, and 20%, respectively. Experiment E4 used the same 10% threshold as E2 but extended the training period back to 1980-09-01.

Table 1 – Regional LSTM experiments with different training periods and missing-data thresholds

Experiment	Full data window	Training range	Missing threshold	No. of catchments	Main comparison
E1	1990–2024	1990-09-01 to 2014-08-31	5%	618	Strict missing-data threshold baseline
E2	1990–2024	1990-09-01 to 2014-08-31	10%	720	Intermediate threshold and reference for period comparison
E3	1990–2024	1990-09-01 to 2014-08-31	20%	801	Permissive missing-data threshold
E4	1980–2024	1980-09-01 to 2014-08-31	10%	668	Effect of adding historical records from 1980–1990

All experiments used the same validation period, from 2014-09-01 to 2019-08-31, and the same test period, from 2019-09-01 to 2024-08-31. Model performance was evaluated using NSE, Kling–Gupta efficiency (KGE), root mean square error (RMSE), mean absolute error (MAE), and percent bias (PBIAS). The comparison among E1–E3 assesses the effect of relaxing the missing-data threshold, while the comparison between E2 and E4 assesses whether adding earlier historical records improves regional model training.

To assess whether missing-data filtering introduced selection bias, the spatial distribution and selected catchment descriptors were compared across missing-data groups. Continuous attributes were examined through distributional analyses, Kolmogorov–Smirnov (KS) tests, and standardized mean differences (SMDs), while regional composition was assessed using relative frequencies within geographic regions.

RESULTS AND CONCLUSION

The missing-data filtering was not spatially or hydroclimatically neutral. Relaxing the threshold from ≤5% to higher missing-data ranges changed both the spatial distribution and the catchment attributes of the retained basins. In the strict ≤5% group, the basins were mainly distributed across the Southeast (35.9%), Northeast (24.1%), and South (21.8%) macro-regions of the Brazilian Institute of Geography and Statistics (IBGE). The basins added under the 5–10% and 10–20% ranges were more concentrated in the Southeast, representing 51.0% and 46.9% of the added basins, respectively (Figure 1). Figure 2 shows that these added basins were also systematically smaller: median drainage area decreased from 3562.7 km² in the strict group to 1325.0 km² in the 5–10% group and 702.0 km² in the 10–20% group. The differences in drainage area were statistically significant according to the

KS test. In addition, the 5–10% group showed wetter and more hydrologically responsive characteristics, including higher mean streamflow, Q95 (95th percentile of daily flow), Q5 (5th percentile), and runoff ratio, along with lower aridity. These results indicate that relaxing the missing-data threshold did not merely increase the number of basins; it also altered the physiographic and hydroclimatic composition of the dataset.

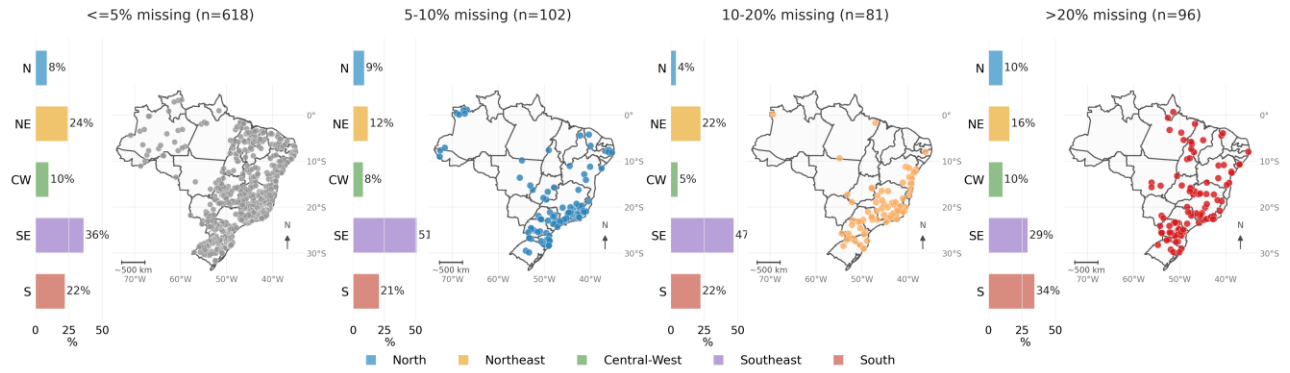


Figure 1. Spatial distribution and Brazilian Institute of Geography and Statistics (IBGE) regional composition of CAMELS-BR basins by streamflow missing-data threshold group

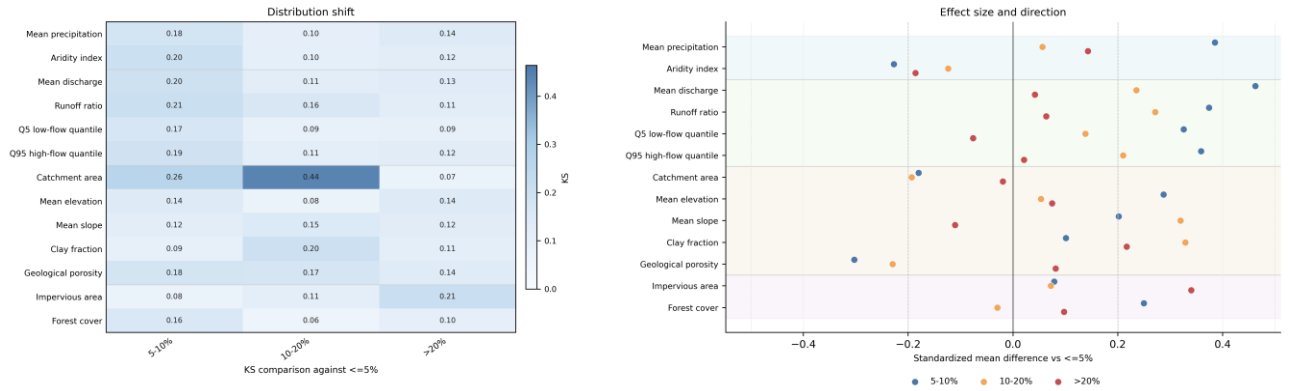


Figure 2. Distributional shift and standardized effect size of selected basin attributes relative to the strict ≤5% missing-data group

Model performance varied across the four experiments (Table 2). The best overall performance was achieved by Experiment 1 (E1), which used the 1990–2024 period and the strict 5% missing-data threshold. Although E1 included the smallest number of basins, it achieved the highest NSE (0.57) and KGE (0.63), and the lowest RMSE (0.74), MAE (0.39), and PBIAS (-0.23%). It also had the highest percentage of basins with NSE > 0.5 and the lowest number of very poor simulations with NSE < -1.

Table 2 – Regional LSTM performance under different training periods and missing-data thresholds

Experiment	Basins	NSE	KGE	RMSE	MAE	PBIAS (%)	NSE > 0 (%)	NSE > 0.5 (%)	NSE < -1 (n)
E1: 1990-2024, 5% missing	618	**0.57**	**0.63**	**0.74**	**0.39**	**-0.23**	93.5	**61.5**	**8**
E2: 1990-2024, 10% missing	720	0.56	0.59	0.77	0.41	-3.11	**95.0**	59.6	11
E3: 1990-2024, 20% missing	801	0.54	0.59	0.78	0.41	-2.36	87.8	54.9	14
E4: 1980-2024, 10% missing	668	0.55	0.58	0.78	0.42	-1.06	93.9	58.8	14

Relaxing the missing-data threshold increased the number of available basins; however, it did not improve model performance. Compared with E1, E2 incorporated 102 additional basins, yet KGE decreased and the bias became more negative. E3 included the largest number of basins, but showed the lowest NSE, fewer basins with NSE > 0.5, and more cases with NSE < -1. Similarly, extending the training period back to 1980 did not improve performance. Compared with E2, E4 used the same 10% threshold but retained fewer basins and produced slightly lower NSE, KGE, RMSE, and MAE, although PBIAS improved.

Overall, the best-performing configuration was not the one with the longest record or the largest number of basins, but the one with the strictest missing-data threshold. This suggests that, for this regional LSTM application in CAMELS-BR, data completeness was more important than maximizing temporal length or catchment sample size. However, because the missing-data threshold also altered the spatial and hydroclimatic composition of the selected basins, performance differences should not be interpreted as the isolated effect of missing-data percentage alone. Instead, they reflect the combined effects of streamflow data quality, catchment availability, and basin representativeness. These results highlight that catchment filtering in CAMELS-BR is not a neutral preprocessing step and should be explicitly reported in regional modeling studies.

REFERENCES

Chagas, V.B.P. *et al.* (2020) "CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil," *Earth System Science Data*, 12(3), pp. 2075–2096. Available at: https://doi.org/10.5194/essd-12-2075-2020.

Heudorfer, B. and Loritz, R. (2026) "Is smart sampling worth it? Impact of training data selection on the performance of LSTMs in streamflow prediction," *Hydrology Research*, 57(3), pp. 501–517. Available at: https://doi.org/10.2166/nh.2026.119.

Kratzert, F. *et al.* (2019) "Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning," *Water Resources Research*, 55(12), pp. 11344–11354. Available at: https://doi.org/10.1029/2019WR026065.

Kratzert, F. *et al.* (2022) "NeuralHydrology — A Python library for Deep Learning research in hydrology," *Journal of Open Source Software*, 7(71), p. 4050. Available at: https://doi.org/10.21105/joss.04050.

Snieder, E. and Khan, U.T. (2025) "A diversity-centric strategy for the selection of spatio-temporal training data for LSTM-based streamflow forecasting," *Hydrology and Earth System Sciences*, 29(3), pp. 785–798. Available at: https://doi.org/10.5194/hess-29-785-2025.

Tripathy, K.P. and Mishra, A.K. (2024) "Deep learning in hydrology and water resources disciplines: concepts, methods, applications, and research directions," *Journal of Hydrology*, 628(October 2023), p. 130458. Available at: https://doi.org/10.1016/j.jhydrol.2023.130458.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the joint support provided by (i) the Hydraulic Laboratory of the School of Engineering, University of São Paulo, and the Hydraulic Technological Center Foundation (FCTH), through funding agreements USP 1019628 and 1021874; and (ii) the São Paulo Research Foundation (FAPESP) through grant No. 2024/02748-7 (E.V.E.S.). The authors also acknowledge the Graduate Program in Civil Engineering of the University of São Paulo (PPGEC-USP) for institutional support during the development of this study.