LSTM BASED SHORT TERM RIVER FORECASTING IN A HIGHLY URBANIZED BASIN USING BAYESIAN HYPERPARAMETER OPTIMIZATION

Thiago M. Osawa [1]; Dário Hachisu Hossoda [2]& Jose Rodolfo S. Martins [3]*

RESUMO – Este estudo avaliou um modelo HydroEcoLSTM baseado em LSTM para previsão do nível da água com antecedência de 120 min na estação de monitoramento, localizada na Bacia do Rio Tamanduateí, São Paulo, Brasil. Foi utilizada uma base hidrometeorológica de 10 min entre 2015 e 2025, com 20 postos pluviométricos selecionados como entradas e o nível do rio como variável-alvo. Os hiperparâmetros foram otimizados com Optuna, utilizando otimização Bayesiana, parada antecipada e poda por mediana. A otimização final avaliou 162 tentativas, com 138 completas e 24 podadas. O melhor modelo individual apresentou RMSE = 0,296 m, MAE = 0,158 m, NSE = 0,798 e KGE = 0,821 no período de teste independente, enquanto o ensemble dos três melhores modelos mostrou desempenho comparável e mais estável. Entretanto, a análise dos resíduos indicou forte autocorrelação temporal, e a avaliação por eventos mostrou limitações na reprodução do tempo de pico, magnitude do pico e volume do hidrograma. Os resultados indicam que modelos LSTM são promissores para previsão urbana de curto prazo, mas requerem melhorias para eventos extremos e sistemas operacionais de alerta.

Palavras-Chave – previsão de inundações urbanas; LSTM; otimização Bayesiana

ABSTRACT–This study evaluated an LSTM-based HydroEcoLSTM model for 120-min river-stage forecasting at a monitoring station in the Tamanduateí River Basin, a highly urbanized and anthropogenically modified catchment in the Metropolitan Region of São Paulo, Brazil. A 10-min hydrometeorological database from 2015 to 2025 was used, including 20 selected rain gauges as dynamic inputs and river-stage observations as the target variable. Hyperparameter tuning was performed with Optuna using Bayesian optimization, early stopping, and median pruning. The final optimization evaluated 162 trials, of which 138 were completed and 24 were pruned, reducing computational cost. The best individual model achieved RMSE = 0.296 m, MAE = 0.158 m, NSE = 0.798, and KGE = 0.821 during the independent test period. The ensemble of the three best models showed comparable performance and lower MAE. However, residual diagnostics revealed strong temporal autocorrelation, and event-based analyses showed limitations in reproducing peak magnitude, peak timing, and hydrograph volume during rainfall-response events. The results indicate that LSTM models are promising for short-term urban river-stage forecasting, but further improvements are needed for extreme-event representation and operational early warning applications.

Key-words – urban flood forecasting; LSTM; Bayesian optimization

1*) Department of Hydraulic and Environmental engineering of University of Sao Paulo, thiago.osawa@usp.br
2) Department of Hydraulic and Environmental engineering of University of Sao Paulo, dario.hossoda@usp.br
3) Department of Hydraulic and Environmental engineering of University of Sao Paulo

INTRODUCTION

Flood forecasting is an essential component of water management and disaster risk reduction, especially in highly urbanized basins where intense rainfall can rapidly generate hazardous hydrological responses. In urban environments, high imperviousness reduces infiltration, increases surface runoff, and shortens the response time between rainfall and river-stage rise. As a result, convective storms and short-duration high-intensity rainfall events can quickly overwhelm drainage systems and produce flash floods, posing risks to infrastructure, mobility, and population safety. In this context, accurate nowcasting and short-term forecasting of river stage or streamflow are critical for early warning systems, real-time control, reservoir operation, evacuation planning, and emergency response. Prolonged or intense rainfall can also saturate urban drainage systems and trigger pluvial flooding, further increasing the need for high-frequency monitoring and forecasting tools

Hydrological models are widely used to support flood forecasting (Byaruhanga et al., 2024), ranging from conceptual and lumped models to distributed physically based approaches. However, the application of these models in dense urban basins can be challenging due to strong anthropogenic modifications, complex drainage networks, channelization, and the influence of detention reservoirs and other hydraulic structures. These factors introduce nonlinearities that are difficult to represent explicitly and may require detailed data and calibration. In this context, data-driven models have become an attractive alternative, as they can provide fast simulations, relatively simple calibration, and lower computational cost while maintaining competitive predictive accuracy. Recent studies have applied machine learning methods such as LinearSVR, XGBoost, and Random Forest to flood forecasting and classification problems, including applications in the Tamanduateí River basin and similar urban environments (Hossoda et al., 2025) (Rocha-filho et al., 2026).

Among data-driven approaches, deep learning models have received increasing attention because of their capacity to learn nonlinear and hierarchical representations directly from time-series data. Compared with traditional machine learning algorithms such as Random Forest, Support Vector Regression, and XGBoost, deep learning models often show stronger predictive performance in data-rich settings, particularly when temporal dependencies are important (Chen, Huang and Huang, 2023). Long Short-Term Memory networks (LSTMs), in particular, are well suited for hydrological forecasting because they can retain information from previous time steps and represent delayed rainfall-runoff responses. Nevertheless, deep learning models may still struggle in highly urbanized and anthropogenically modified basins and short-term monitoring data available (An and Ouyang, 2025).

Despite the increasing use of machine learning in flood forecasting, two gaps remain relevant for the Metropolitan Region of São Paulo (MRSP). First, few studies have evaluated deep learning models for short-term river-stage forecasting in this region, particularly in highly urbanized basins such as the Tamanduateí. Second, systematic hyperparameter tuning of deep learning models is still limited in many hydrological applications, even though model architecture, sequence length, learning rate, dropout, and batch size can strongly affect forecasting performance. Addressing these gaps is important for developing robust models that can support operational early warning systems under local hydrometeorological conditions.

Therefore, this study aims to develop and evaluate an LSTM-based HydroEcoLSTM model for short-term river-stage forecasting in the Tamanduateí River basin, a highly urbanized and anthropogenically modified catchment in the MRSP. The model is calibrated using high-frequency rainfall and river-stage observations, and Bayesian optimization is applied to identify suitable hyperparameter configurations. The study evaluates model performance over an independent test period using general deterministic metrics and event-based analyses, with particular attention to rainfall-response events, peak timing, peak magnitude, and hydrograph volume. The results provide a benchmark for future deep learning applications in urban flood forecasting and contribute to the development of decision-support tools for early warning systems.

STUDY SITE AND DATABASE

This study focuses on the Tamanduateí River Basin (TRB), located in the MRSP, Brazil (Figure 1). The basin covers approximately 330 km² and drains six highly urbanized municipalities: São Paulo, São Bernardo do Campo, Diadema, Santo André, Mauá, and São Caetano do Sul. The main river extends for 36.5 km and has been extensively channelized and rectified, especially along the Avenida dos Estados, a key urban infrastructure corridor in the region. The basin is characterized by strong urbanization and high soil imperviousness, estimated at 88.2%, which results in rapid rainfall-runoff response and frequent flash-flood susceptibility (de Mira *et al.*, 2026)

The study area has also been strongly modified by flood-control infrastructure. Since 1999, 21 detention ponds, locally known as *piscinões*, have been implemented upstream, with a total design storage capacity approximately of approximately 5.7 hm³. These structures alter the natural rainfall-runoff response and introduce additional non-linearity into the hydrological system. Although the theoretical concentration time is estimated at 230 min, reservoir operation and urban drainage controls can modify the effective hydraulic response time (de Mira *et al.*, 2026).

The target monitoring point is station 413, located near the Municipal Market along the Tamanduateí River. The operational alert thresholds adopted for this station are: attention level at 722.779 m, warning level at 723.779 m, emergency level at 724.779 m, and overflowing level at 725.779 m.

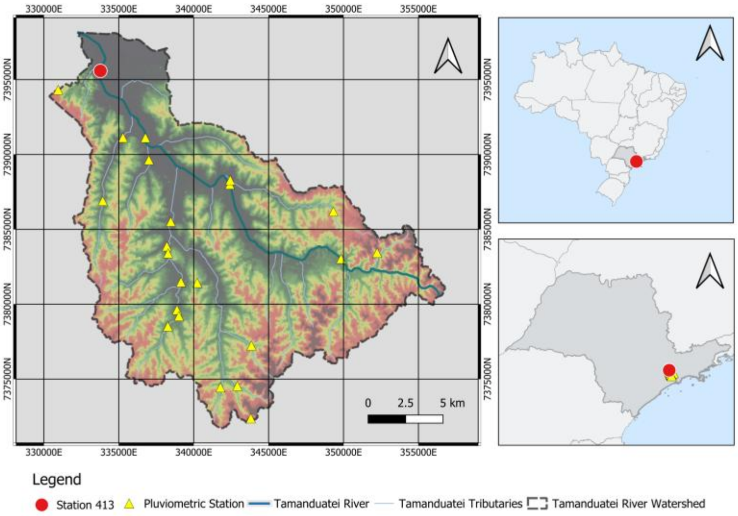

Figure 1 – Localization Map of Tamanduateí River Watershed

The hydrometeorological database was obtained from the São Paulo Flood Alert System (SAISP) operated by the School of Engineering Hydraulic Laboratory and covers the period from 1 April 2015 to 29 March 2025 at 10-min temporal resolution. The dataset includes 23 rainfall gauges as dynamic input features and river-stage observations from station 413 as the target variable. Following (Escobar-Silva *et al.*, 2025), the raw time series were pre-processed using a Hampel outlier detector and a zero-phase low-pass filter to reduce noise while preserving the timing of flood peaks.

METHODOLOGY

This study used the HydroEcoLSTM framework (Nguyen et al., 2025) for short-term river-stage forecasting at station 413. The workflow was adapted to support long hyperparameter optimization using Optuna, including SQLite-based study persistence, resumable optimization, early stopping, and median pruning of low-performing trials. These additions allowed the optimization process to be interrupted and resumed without losing previous results, while reducing computational cost.

The hydrometeorological dataset was chronologically divided into training, validation, and testing periods using a 60/20/20 temporal split (Pravin et all, 2022). This partition was adopted to reproduce operational forecasting conditions, in which the model is evaluated on future unseen events. Although datasets with different lead times were prepared during the exploratory stage, the final optimization and performance analyses presented in this study were conducted for a 120-min forecast horizon, using stage data from the considered monitoring station as the target variable.

Hyperparameter tuning was performed using Bayesian optimization with Optuna. The search space included sequence length, learning rate, hidden size, batch size, dropout rate, and number of LSTM layers. Optuna's Tree-structured Parzen Estimator sampler was used to propose new configurations based on previous trial performance, while a median pruning strategy interrupted underperforming trials after a warm-up period. The optimization was organized in sequential phases, starting with broader exploratory searches and ending with a refined final calibration using longer training and extended patience.

The LSTM model was trained by minimizing Root Mean Squared Error (RMSE) loss over the training data, while validation loss was used to guide hyperparameter selection and early stopping. After optimization, the best hyperparameter configurations were selected according to validation loss and independently evaluated over the unseen test period. Model performance was quantified using RMSE, Mean Absolute Error (MAE), Nash-Sutcliffe Efficiency (NSE), Kling-Gupta Efficiency (KGE), and Percent Bias (PBIAS). The use of those errors indicators are not general when related to hydrological data (Hodson, 2022). Additional analyses were conducted for high-flow and low-flow conditions, as well as event-based rainfall-response windows, including peak timing error, peak magnitude error, and hydrograph volume difference.

To assess robustness, the best individual configurations were compared with an ensemble prediction obtained by averaging the simulated values of the selected top-performing models at each time step. Since the model target represents the river stage 120 min ahead, event-based comparisons were aligned by shifting model predictions from forecast issue time t to target time t + 120 min.

RESULTS

Input dependence among rainfall gauges was evaluated using Pearson correlation and Kendall's tau. Highly dependent stations were removed to reduce redundancy and potential overfitting. Stations 279, 280, and 1000350 were excluded due to strong dependence when compared with nearby gauges data, with Kendall's tau values of 0.67, 0.76, and 1.00, and coefficients of determination of 0.77, 0.80, and 1.00, respectively. After this filtering step, 20 rainfall gauges remained as model input variables.

Hyperparameter Tuning

A refined hyperparameter search was conducted using Optuna to calibrate the HydroEcoLSTM model for station 413. Based on the previous optimization phases, the final search space was restricted to sequence lengths between 500 and 1000 timesteps, learning rates from 0.0005 to 0.0008, hidden sizes from 128 to 256 units, batch sizes from 64 to 88, dropout rates from 0.2 to 0.6, and one or two LSTM layers. Each trial was trained for up to 100 epochs, with early stopping patience of 30 epochs, and the validation loss was used as the optimization objective. The best validation-loss configurations identified during this phase are summarized in Table 1.

Table 1 – Hyperparameters combination of the best validation loss model during optimization

rank	ID	seq_len	lr	hidden	batch	dropout	Layers	Val Loss
1	49	644	0.0007	160	88	0.6	2	0.021
2	3	644	0.0007	128	88	0.4	2	0.022
3	50	644	0.0007	160	88	0.6	2	0.022

The Phase 4 optimization evaluated 162 trials, of which 138 were completed and 24 were pruned. The full run required approximately 35.2 hours. Completed trials ran for a median of 88.5 epochs, whereas pruned trials were interrupted after a median of 42 epochs. By stopping low-potential configurations after the warm-up period, the pruning strategy avoided approximately 1328 additional epochs, corresponding to an estimated computational saving of 3.4 to 3.7 hours. The distribution of validation losses obtained during the optimization is shown in Figure 2.

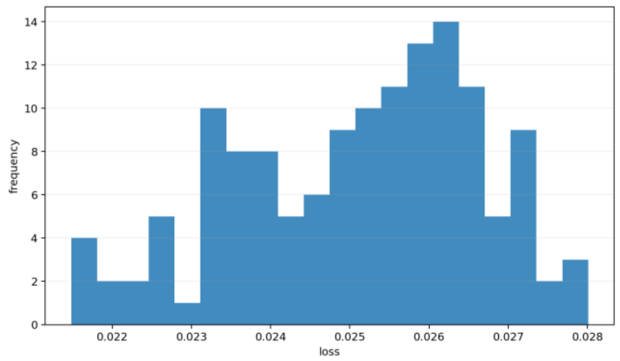

Figure 2 – Loss distribution frequency in optimization

Overall, the Optuna-based strategy provided a more efficient alternative to exhaustive grid search and purely random search by adaptively guiding the exploration toward promising regions of the hyperparameter space. The best validation-loss configuration was obtained in trial 49, with a validation loss of 0.02148, followed closely by trials 6 and 50. The similarity among the best configurations, particularly in sequence length, learning rate, batch size, and number of layers, suggests convergence toward a consistent region of the hyperparameter space.

Model Efficiency

Among the evaluated models, the best individual LSTM configuration was rank_03_trial_0050, which achieved the lowest RMSE and the highest NSE among the trained models, with RMSE = 0.296m and NSE = 0.798 (Table 2) during the independent test period. This indicates that this configuration provided the best overall reproduction of the observed water level dynamics, particularly when larger deviations are penalized. The mean ensemble of the three best models showed comparable performance, with RMSE = 0.298 m and NSE = 0.795, and achieved the lowest MAE among the LSTM-based approaches, suggesting that ensemble averaging reduced part of the individual model variability. Therefore, considering the objective of forecasting dynamic river-level responses, the rank_03_trial_0050 model can be considered the best individual model, while the ensemble represents a robust alternative with more stable average errors.

Table 2 – Summary of model's performance metrics

model	RMSE	MAE	NSE	KGE	PBIAS	FHV%	FLV%
rank_01_trial_0049	0.305	0.153	0.785	0.788	-0.072	-0.132	**0.002**
rank_02_trial_0006	0.310	0.153	0.779	0.757	**-0.007**	-0.137	0.004
rank_03_trial_0050	**0.296**	0.158	**0.798**	**0.821**	0.046	**-0.099**	0.004
top3_mean_ensemble	0.298	**0.147**	0.795	0.789	-0.011	-0.123	0.003

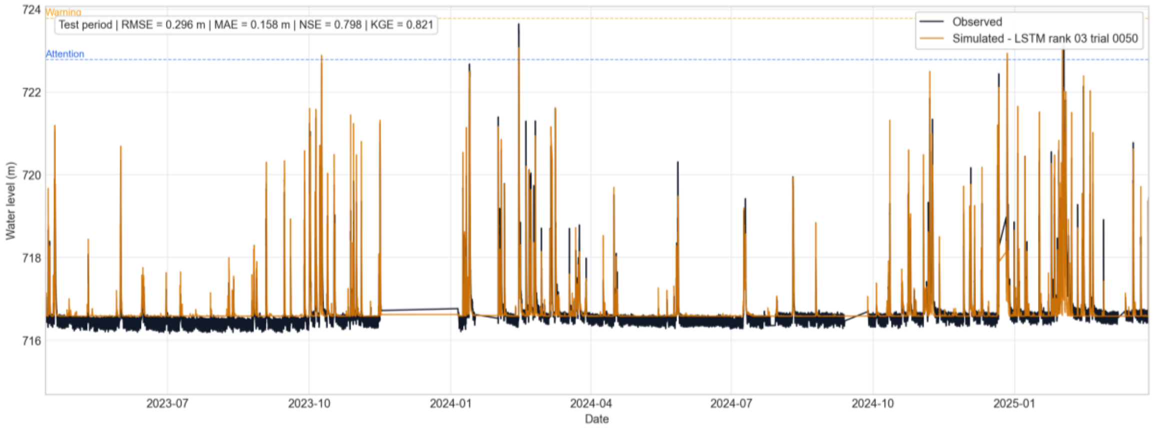

Figure 3 – Observed and simulated river stage for the rank 3 model, trial 50, during the independent test period

The autocorrelation function (ACF) and partial autocorrelation function (PACF) were used to evaluate whether model residuals were temporally independent. While the ACF measures the correlation between residuals and their past values at different lags, the PACF estimates the direct correlation at each lag after removing the influence of shorter lags. For an ideal model, residuals should approximate white noise, with ACF and PACF values close to zero.

Figure 4 shows that the residuals of the selected model still presented strong temporal autocorrelation during the test period. The ACF showed a lag-1 correlation of 0.972 at the 10-min time step, far above the 95% white-noise confidence bounds, and the correlation remained positive over several hours, with values of 0.568 at 1h, 0.364 at 2h, and 0.247 at 6h. The PACF was estimated on a stable 30-min sample, also indicated strong short-term dependence, with a lag-1 partial autocorrelation of 0.822. Together with the low Durbin-Watson statistic of 0.055, these results indicate that the residuals are not white noise and that model errors persist through time. Therefore, although the model achieved satisfactory deterministic performance, part of the temporal structure of the river-stage process remained unexplained, representing an important limitation for event-scale forecasting and early warning applications.

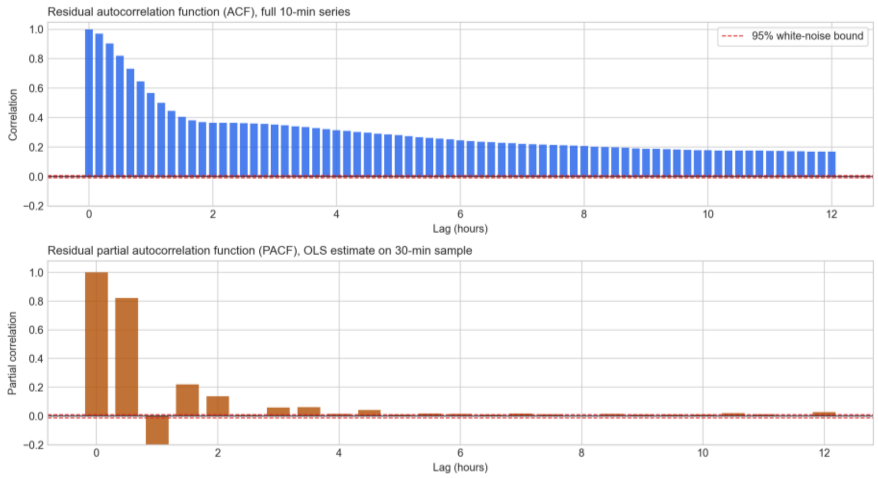

Figure 4 – Residual autocorrelation diagnostics for the selected LSTM model during the test period

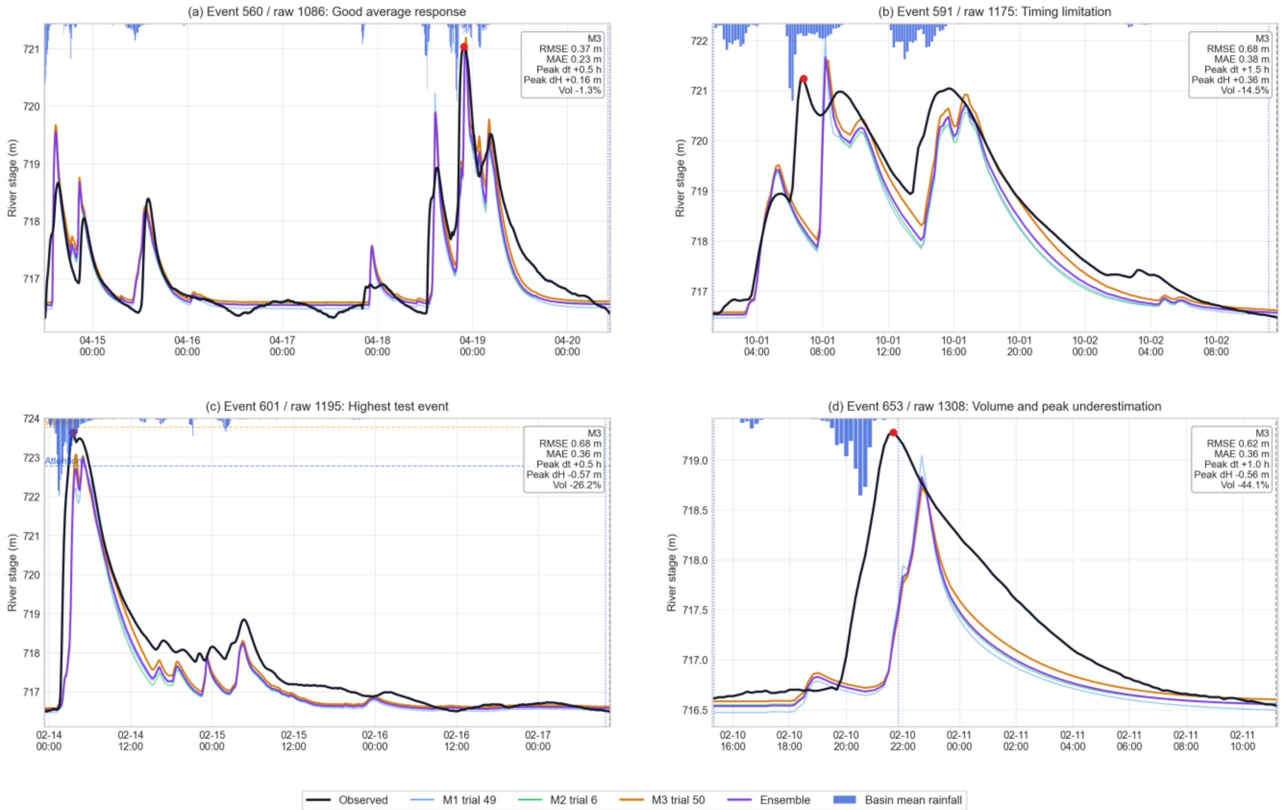

Figure 5 – Event-based comparison between observed and simulated hydrographs during selected rainfall-response events in the test period

The selected event hydrographs (Figure 5) show that the model was able to reproduce the overall timing and shape of moderate rainfall-response events, as illustrated by event 560. However, performance decreased during more challenging events, particularly those associated with higher river stages or sharper hydrograph responses. Event 601, the highest event in the test period, shows that the model underestimated peak magnitude and event volume, while event 653 highlights a stronger underestimation of hydrograph volume. Event 591 further illustrates that even when the general response pattern is captured, errors in peak timing may still occur. These results indicate that although the model performs well under average conditions, limitations remain in representing event-scale dynamics, especially peak magnitude, timing, and hydrograph volume.

CONCLUSION, LIMITATIONS AND FUTURE WORKS

This study evaluated an LSTM-based HydroEcoLSTM model for short-term river-stage forecasting at station 413 in the Tamanduateí River basin. The results show that the proposed framework is promising for high-frequency forecasting and potential early warning applications. The model achieved relatively low RMSE and MAE values, indicating small absolute errors in water-level units. However, NSE and KGE were more moderate, suggesting that the model did not fully reproduce the temporal variability, amplitude, and correlation structure of the observed series. This indicates that low average errors may coexist with limitations in representing event dynamics, peak timing, hydrograph volume, and peak magnitude.

The Bayesian/Optuna optimization strategy was effective for exploring the hyperparameter space while reducing computational cost through pruning. However, several configurations produced similar validation performance, showing that selecting a single model based only on validation loss may not be ideal. The best validation-loss model did not necessarily provide the best performance on the independent test period. This highlights the importance of evaluating multiple candidate models using complementary metrics. The ensemble model, based on the average of the best configurations, did not always produce the best individual metric values, but showed consistent performance across different criteria, making ensemble approaches a promising direction for improving model robustness.

A key limitation was the model's reduced ability to represent extreme hydrological responses. The model performed better under average river-stage conditions but showed limitations in reproducing peak magnitude, peak timing, and event hydrograph volume. This is likely related to the scarcity of extreme rainfall and high-stage events in the training data. Future studies should evaluate weighted loss functions, event-based training objectives, and data augmentation techniques to increase the importance of rare but operationally relevant events during calibration.

Residual diagnostics also indicated temporal autocorrelation, suggesting that model errors were not fully random. This may indicate that the model partly reproduced river-stage persistence rather than fully learning the rainfall-runoff response of the basin. In addition, strong dependence among rainfall gauges may have reduced the effective spatial information available to the model. Future work should investigate feature selection, rainfall station clustering, station-specific lag analysis, and input reduction methods to better represent the spatial and temporal structure of rainfall.

Another source of uncertainty is the random initialization of model weights and biases. Different random seeds can lead to different trained models, even for the same hyperparameter configuration. Future studies should evaluate repeated runs with multiple seeds or report uncertainty associated with model initialization.

Finally, average performance metrics alone are insufficient for early warning applications. Future evaluations should include event-based and categorical metrics, such as peak timing error, peak magnitude error, hydrograph volume difference, false positives, false negatives, and threshold-based accuracy for attention, warning, emergency, and overflowing classes. This work provides a benchmark for future LSTM-based short-term river-stage forecasting in the Tamanduateí River basin, while highlighting the need to improve extreme-event representation, reduce residual autocorrelation, and account for anthropogenic controls such as reservoir and drainage system operations.

ACKNOWLEDGEMENTS - The authors gratefully acknowledge the support provided by the Hydraulic Laboratory of the School of Engineering, University of São Paulo, and by the Hydraulic Technological Center Foundation (FCTH), through funding agreements USP 1019628 and 1021874. 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.

REFERENCES

An, H. and Ouyang, C. (2025) "A hybrid framework for real-time flash flood forecasting in small ungauged catchments: integrating hydrodynamic simulations with LSTM networks," *Journal of Hydrology*, 661. Available at: https://doi.org/10.1016/j.jhydrol.2025.133688.

BYARUHANGA, Nicholas et al. (2024). " Evolution of flood prediction and forecasting models for flood early warning systems: A scoping review,". Water, v. 16, n. 13, p. 1763.

Chen, S., Huang, J. and Huang, J.C. (2023) "Improving daily streamflow simulations for data-scarce watersheds using the coupled SWAT-LSTM approach," *Journal of Hydrology*, 622. Available at: https://doi.org/10.1016/j.jhydrol.2023.129734.

Escobar-Silva, E. V. *et al.* (2025) "TTI-HydroMet: A Decade of High-Resolution Rainfall and Streamflow for the Tamanduateí River Watershed, Brazil."

Gouveia, I.C.M.C. and Rodrigues, C. (2025) "The importance of retrospective geomorphological cartography for the definition of complex morphological units – The Tamanduatei river basin, metropolitan region of Sao Paulo". *Derbyana*, 46. Available at: https://doi.org/10.69469/derb.v46.835.

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

Hossoda, D.H. *et al.* (2025) "Data-Driven Modeling for Urban Flood Warning Systems: A Case Study in the Guarará Basin, Brazil," *Journal of Flood Risk Management*, 18(3), pp. 1–13. Available at: https://doi.org/10.1111/jfr3.70110.

Hodson, T. O.(2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not, Geosci. Model Dev., 15, 5481–5487, https://doi.org/10.5194/gmd-15-5481-2022.

de Mira, Í.R.C. *et al.* (2026) "Flooding in the Tamanduateí River Hydrographic Basin: A Spatial and Multifactorial Analysis," *Transactions in GIS*, 30(1). Available at: https://doi.org/10.1111/tgis.70213.

Nguyen, T. V. *et al.* (2025) "HydroEcoLSTM: A Python package with graphical user interface for hydro-ecological modeling with long short-term memory neural network," *Ecological Informatics*, 85(January), p. 102994. Available at: https://doi.org/10.1016/j.ecoinf.2025.102994.

PRAVIN, P. S. et al. (2022). Hyperparameter optimization strategies for machine learning-based stochastic energy efficient scheduling in cyber-physical production systems. Digital Chemical Engineering, v. 4, p. 100047.

Priya, R.Y. and Manjula, R. (2025) "Enhancing daily streamflow prediction: A comparative analysis of univariate LSTM and N-BEATS models with coupled SWAT-LSTM and SWAT-N-BEATS models incorporating influential SWAT features," *Journal of Earth System Science*, 134(2). Available at: https://doi.org/10.1007/s12040-025-02567-5.

Rocha-filho, K.L. *et al.* (2026) "Ten years of hydrometeorological observations at 10-minute resolution and its application in machine learning hydrological models," (January), pp. 1–20.