WAT-50 ORAL PRESENTATION · WATER TRACK · SURFACE WATER HYDROLOGY & BASIN MANAGEMENT

Data Quality as Part of the Model Performance: Benchmarking Missing-Value Imputation for Tropical Streamflow Prediction

A. H. Z. Aqthaf[1*] and R. L. H. L. Rajapakse[1]

[1]*Department of Civil Engineering, Faculty of Engineering, University of Moratuwa, Katubedda, Sri Lanka*

**Corresponding author: aqthaf.hm@gmail.com*

DOI: https://doi.org/10.31705/ICEWSE.2026.66

Abstract

Globally, reliable hydrological prediction depends not only on model architecture but also on how missing hydro-meteorological observations are reconstructed before model training. This issue is critical in data-scarce tropical basins, where missing rainfall, temperature, evaporation, and streamflow observations are common, and imputation methods are often selected by convention rather than by evidence. This research benchmarks twelve missing-value imputation methods across four Sri Lankan river basins, four variable classes, and five synthetic missingness levels (5–40%) generated as short (1–3-day) and occasional longer (20–30-day) dropouts mimicking sensor failure, yielding 960 completed method–basin–variable–missingness evaluations. The methods span linear interpolation, Kalman filtering, k-nearest neighbours, multiple imputation by chained equations, missForest, Gaussian-copula and gradient-boosted approaches, together with self-attention and bidirectional recurrent deep imputers. A Friedman test demonstrates significant performance differences among methods ($p<0.001$), with SAITS and BRITS achieving the strongest ranks. However, linear interpolation remains statistically competitive with several established methods, matches SAITS for streamflow reconstruction, and requires approximately 18,000 times less computation. Performance depends strongly on variable and basin characteristics; notably, an STL-plus-spline approach fails severely for bursty tropical rainfall. When reconstructed inputs are propagated through a downstream XGBoost streamflow model, the mean spread in validation Nash–Sutcliffe efficiency across imputation methods is only 0.091, while 74.5% of all method–basin–missingness combinations exceed NSE values of 0.80. These results show that imputation accuracy does not translate into proportionate gains in hydrological prediction. The study therefore proposes a performance- and application-based decision framework in which imputation sophistication is matched to downstream sensitivity, data characteristics, and computational constraints.

Keywords: *computational efficiency; decision framework; deep learning; gap-filling; hydrological time series; Sri Lankan river basins*