

Integration of Generated Weather Data into LSTM-Training to Improve the Simulation of Extreme Flood Events

Paul Reis¹, Alexander Dolich², Antara Dasgupta³, Paul Hassenjürgen¹, Sergiy Vorogushyn⁴, Viet Dung Nguyen⁴, and Ralf Loritz²

¹RWTH Aachen University, Institute of Hydraulic Engineering and Water Resources Management, Aachen, Germany (paul.reis@rwth-aachen.de)

²Karlsruhe Institute of Technology, Institute of Water and Environment, Chair of Hydrology, Karlsruhe, Germany

³Junior Professor for Data Driven Computing in Civil Engineering, RWTH Aachen University

⁴GFZ Helmholtz Centre for Geosciences, Section Hydrology, Potsdam, Germany

1. Problem Statement

- HBV (conceptual model) and benchmark-LSTM (Long Short-Term Memory Network, data driven model) systematically underestimate peak discharge, worsening with increasing Average Recurrence Interval (ARI)
- Benchmark-LSTM saturates with artificial 1-day event at 43.5 mm/d, below theoretical limit of 53.3 mm/d
- HBV extrapolates linearly beyond observed precipitation ranges but degrades most for extreme floods above ARI>25a

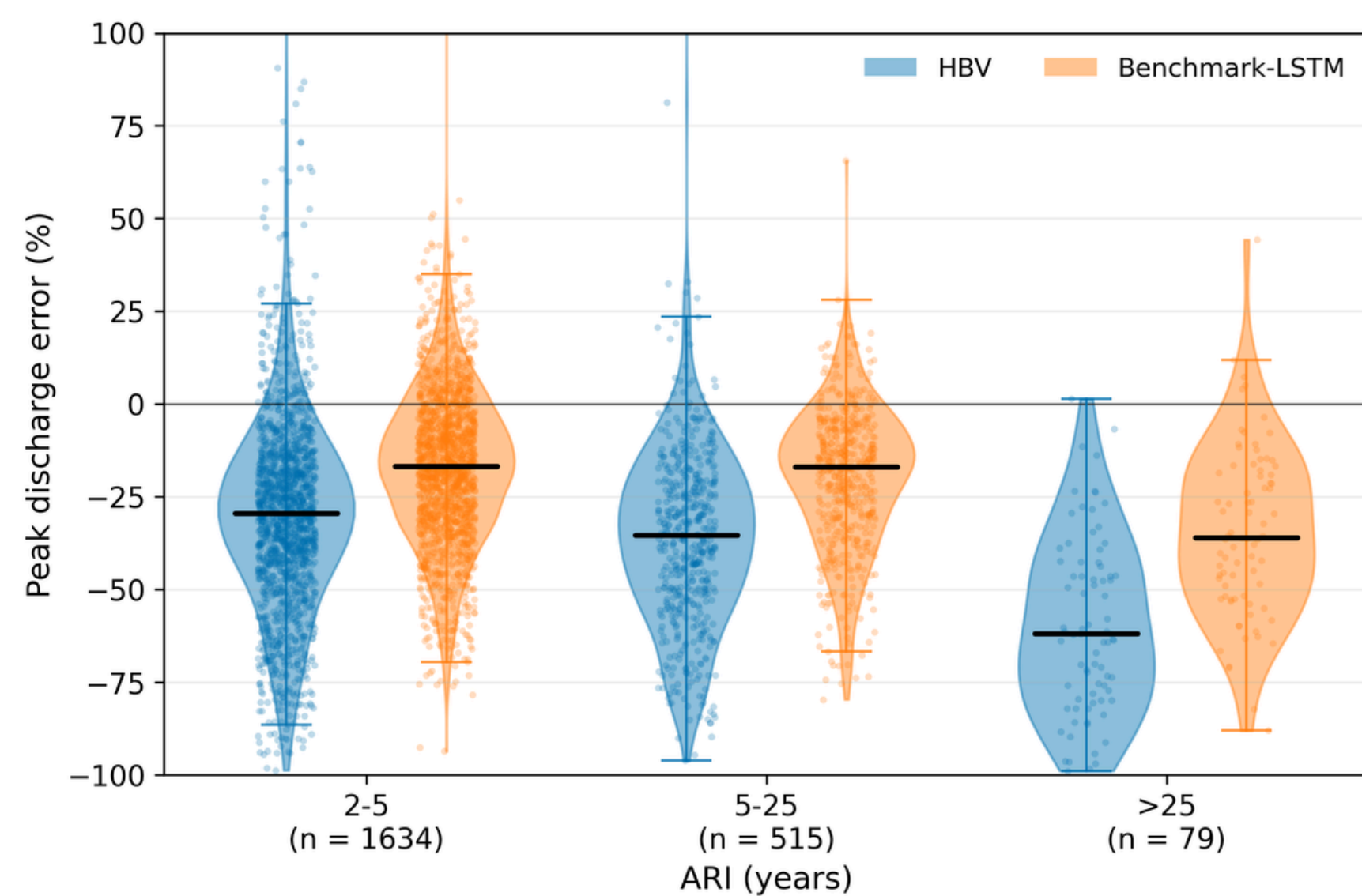


Fig. 1: Peak discharge error (%) by return period range for CAMELS-DE catchments over the testing period. Negative values indicate underestimation. Violin plots show the distribution of peak discharge errors, dots represent individual catchments. Black bars indicate the median, colored whiskers span the 1.5*IQR range.

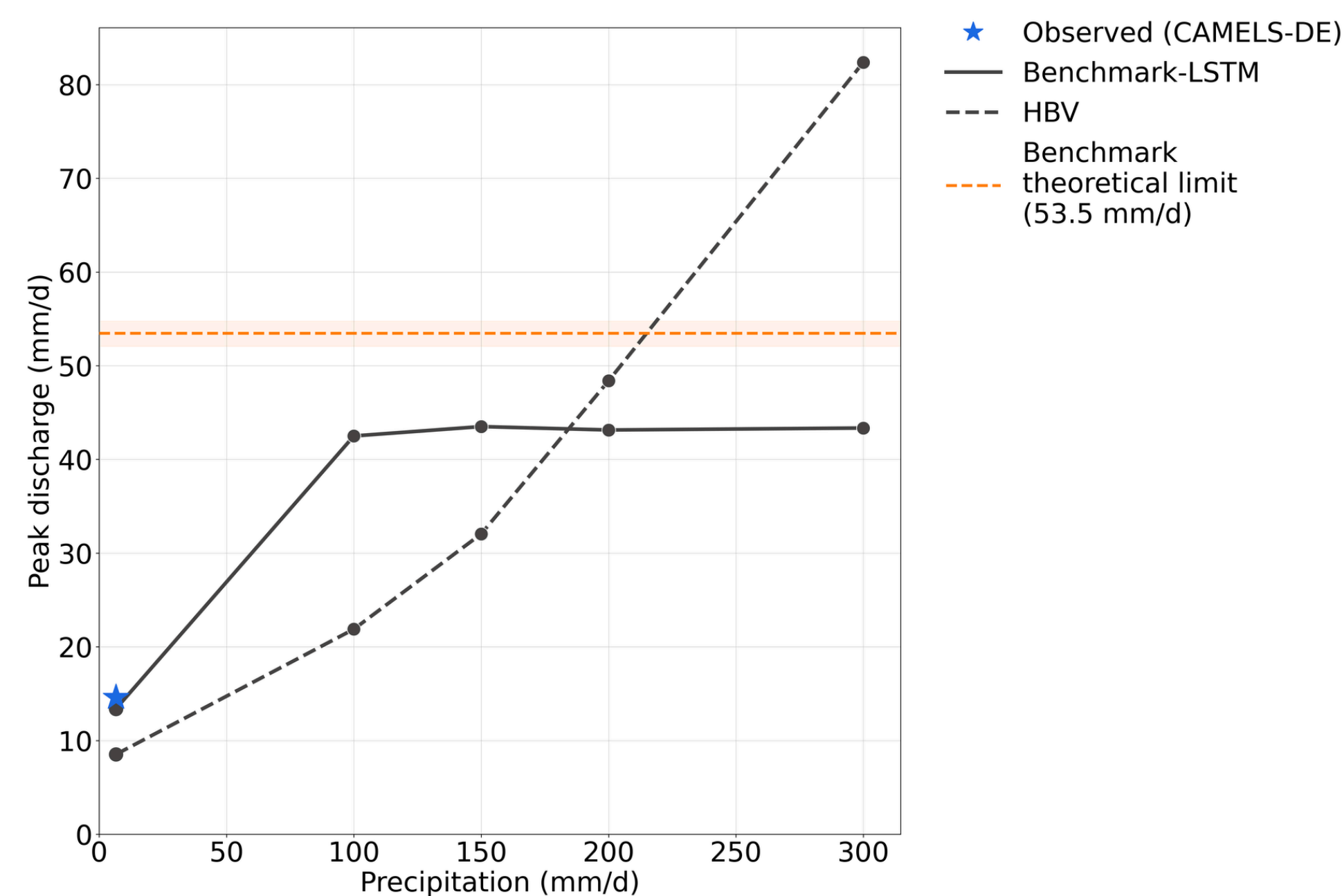


Fig. 2: Simulated peak discharge for the observed max. event and 4 artificial 1-day precipitation scenarios in Müsch (DEB11230), demonstration of LSTM saturation and HBV extrapolation capabilities.

2. Datasets

- CAMELS-DE: 70 years observed data
- GFZ Weather Generator (nsRWG): 100 x 72 years synthetic data
- 169 lumped catchments in NRW, Germany

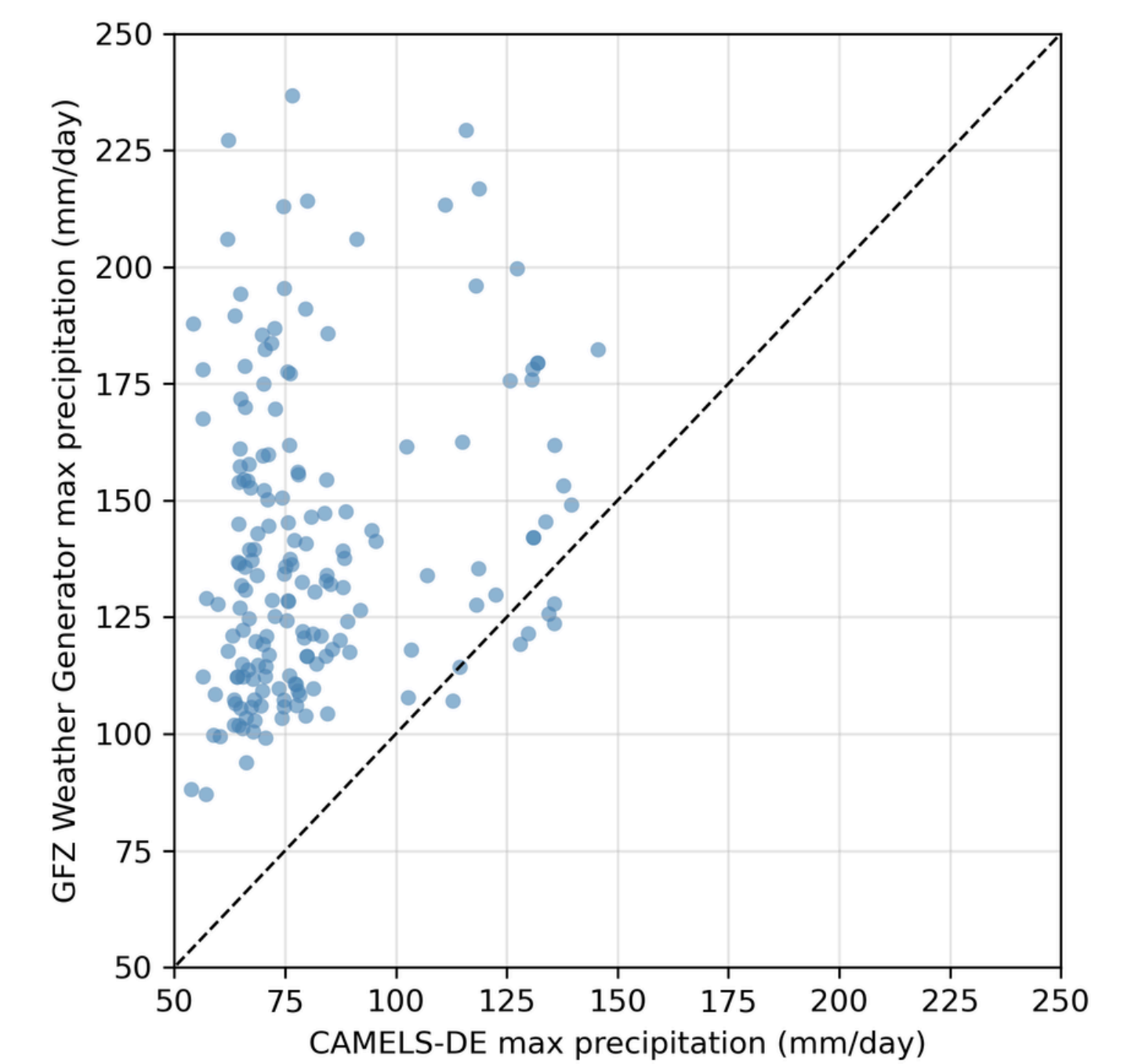


Fig. 3: Max daily precipitation per catchment, observed CAMELS-DE vs. GFZ non-stationary Regional Weather Generator.

3. Data Augmentation Process

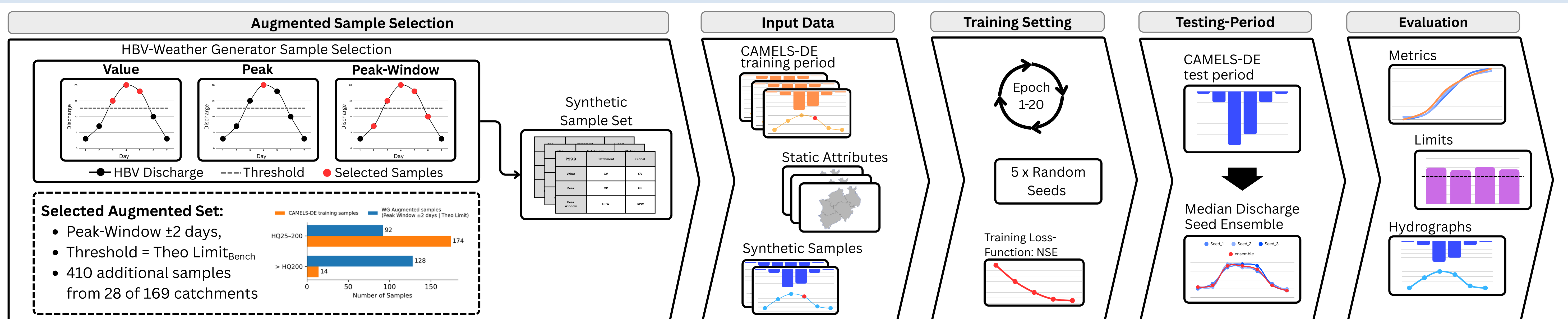


Fig. 4: Data augmentation process diagram

4. Limit Evaluation

- To evaluate practical limit an artificial 1-day precipitation event was used
- Theoretical discharge limit raised by +29% compared to benchmark-LSTM
- Practical limit improved across all NRW catchments, in median by +45% compared to benchmark-LSTM

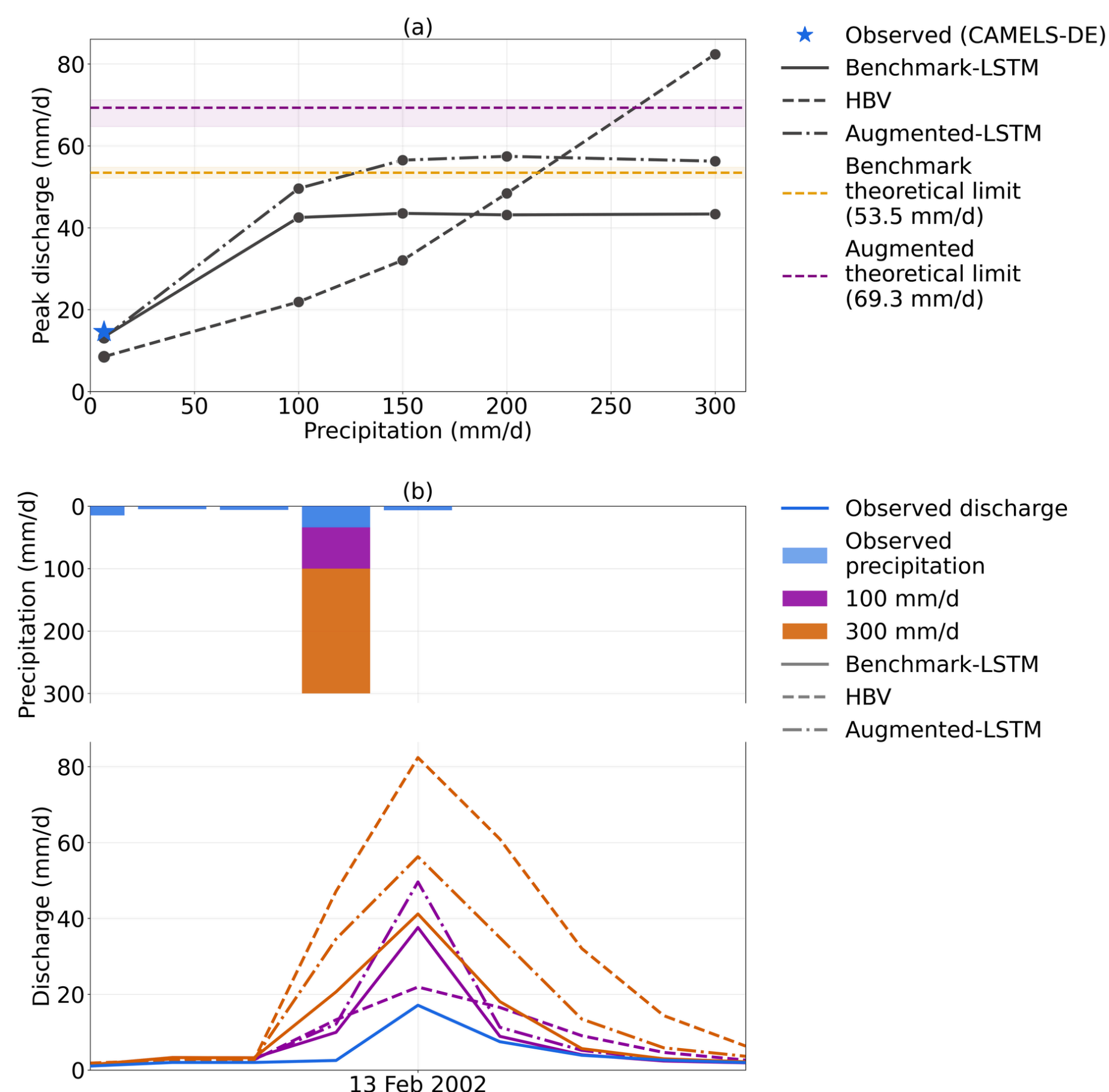


Fig. 5: Simulation evaluation for artificial 1-day event, in Müsch (DEB11230) with data augmentation approach: Peak Window ± 2 days, Theo Limit threshold: (a) Increase of theoretical limit, LSTM_{aug} saturates at higher precipitation. (b) Hydrograph demonstrating LSTM prediction limit. Figure based on Baste et al., 2025

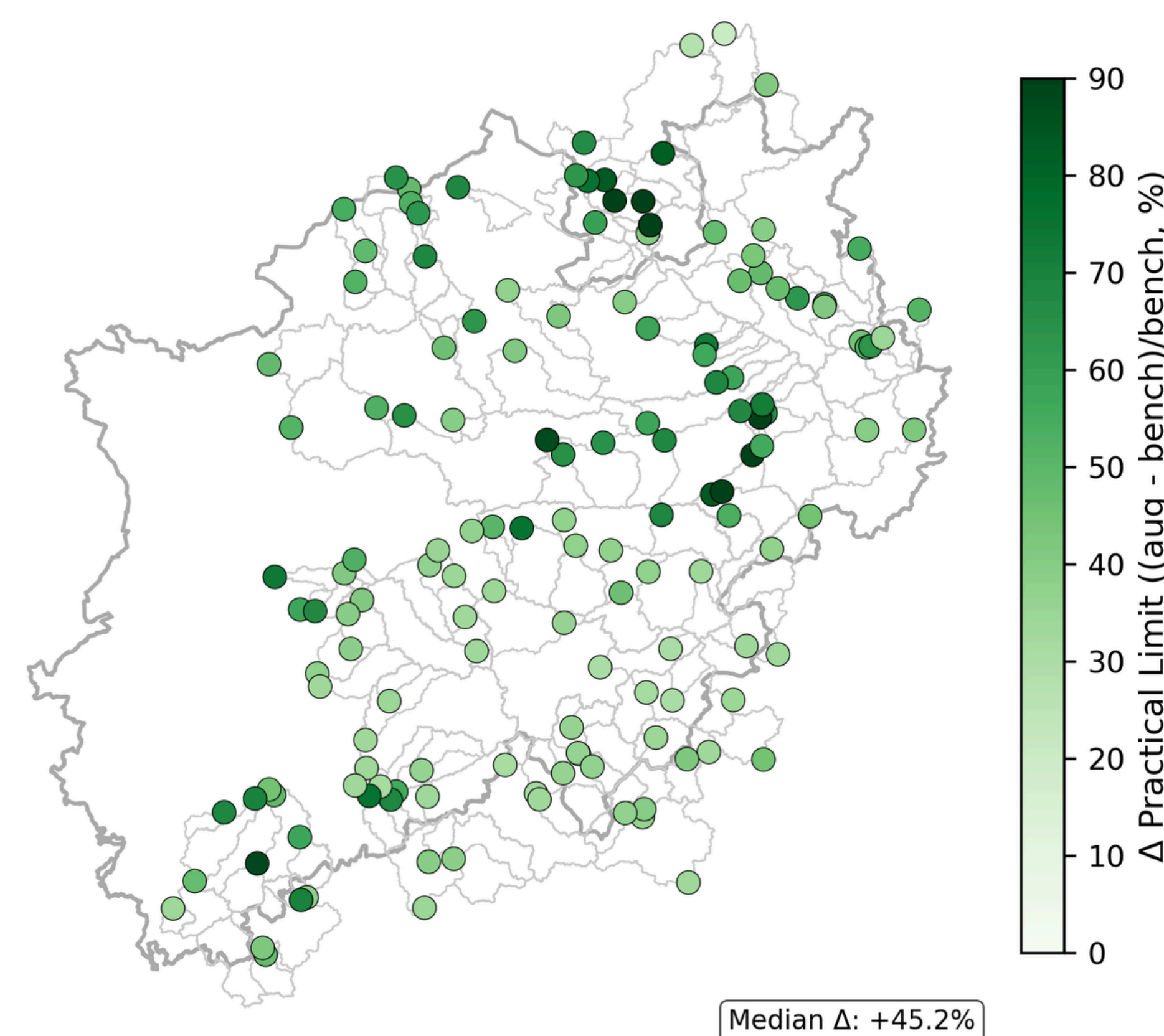


Fig. 6: Per-catchment improvement of practical discharge limit using augmented-LSTM, augmentation approach: Peak Window ± 2 days, Theo Limit threshold over benchmark-LSTM for 169 catchments in research area.

5. Performance Evaluation

- Overall LSTM-model quality preserved:
 - Median NSE (Nash-Sutcliffe Efficiency) range for augmented-LSTM sets: 0.87 - 0.90
- All LSTM models outperform HBV
 - Median HBV NSE: 0.78

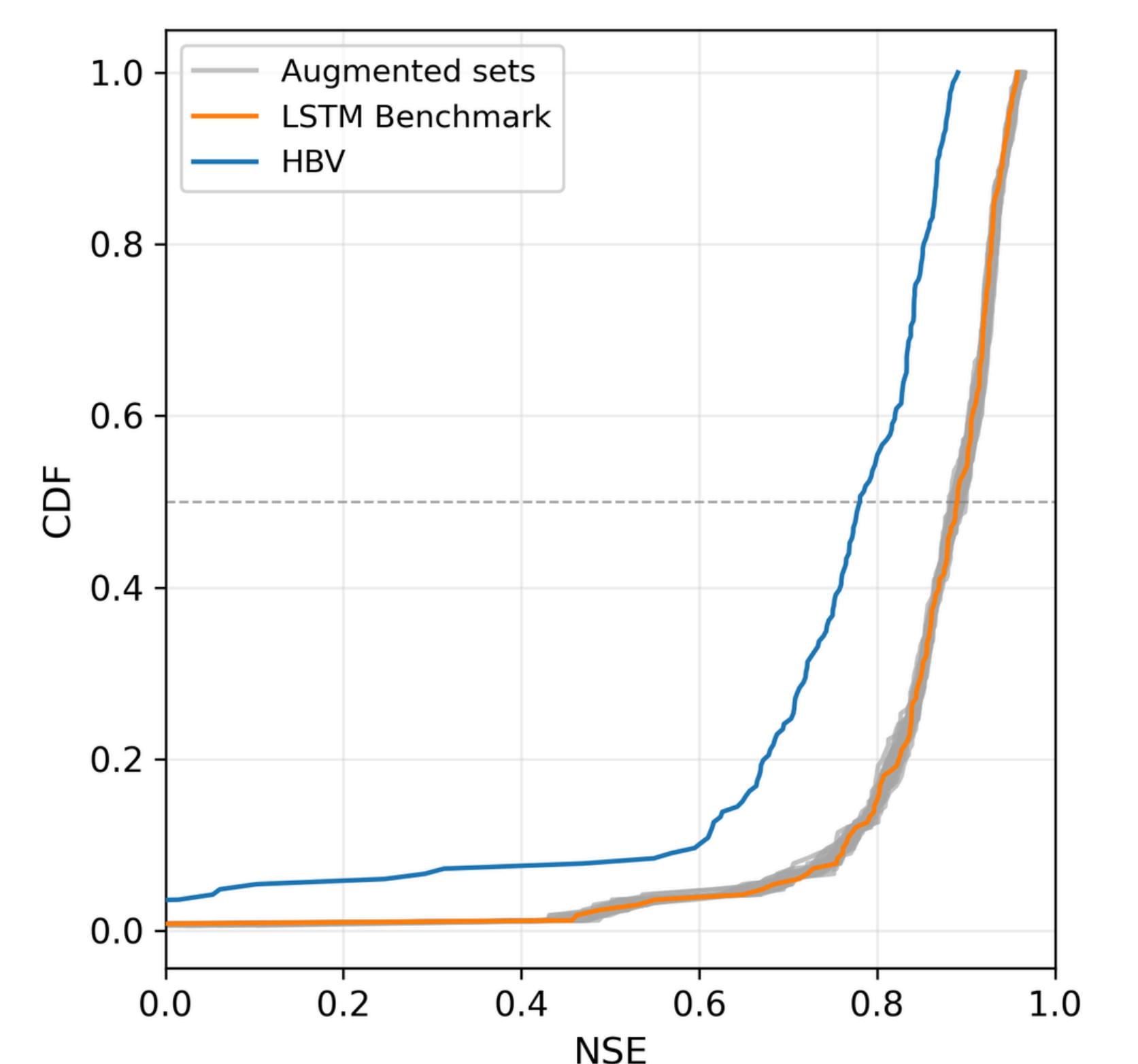


Fig. 7: Cumulative Distribution Function (CDF) of NSE for 169 catchments in research area for test period. Median NSE for bench-LSTM is 0.89, for HBV is 0.78 and for the augmented-LSTM-sets ranging from 0.87 - 0.90.

6. Key-Findings

- NSE stable (0.87–0.90), augmentation preserves overall performance
- Theoretical limit +29%, practical limit +45% across NRW
- Peak flow improves, rarest events (ARI>100) benefit most
- No substantial low-flow penalty

7. Outlook

- Improve HBV extreme-event performance to boost augmentation sample quality
- Investigate theoretical limit drivers: frequency vs. magnitude
- Evaluate peak error reduction for rarest observed events

