

Summary:

The manuscript, “Streamflow prediction in data-scarce regions with semi-supervised deep learning,” proposes a two-stage semi-supervised learning method for streamflow prediction in data-scarce regions. This two-stage framework first involves pre-training an encoder and an LSTM network on large volumes of “unlabeled” data (i.e., meteorological forcings and basin attributes) using Contrastive Predictive Coding (CPC). Then, it involves fine-tuning the pre-trained network on limited amount of “labeled” data (i.e., meteorological forcings and streamflow measurements) using conventional supervised learning methods. This work evaluates the proposed framework for both regional and single-basin models under in-domain and out-of-domain prediction settings, and compares its performance with that of a conventional supervised LSTM model. The comprehensive performance analysis, based on a range of evaluation metrics—including those targeting low-flow and high-flow conditions—provides valuable insights into the strengths of the proposed approach.

Overall, this study provides a meaningful and solid contribution to the development of streamflow prediction methods for data-scarce basins, where sufficient streamflow observations for supervised learning models are often unavailable or of limited quality. I therefore recommend the manuscript for publication, subject to minor revisions.

Response: We thank the reviewer for thoughtful comments and for taking the time to review our work.

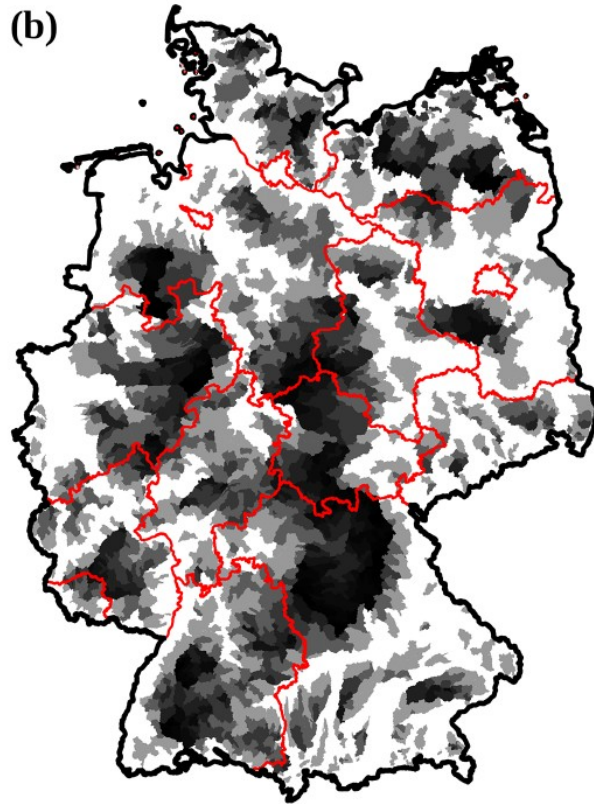
Major comments:

1. The fine-tuning datasets (e.g., Train80%) were generated by randomly sampling from the full dataset. Could the authors explain why this strategy was chosen instead of using temporally continuous subsets (e.g., the first 30, 20, 10, or 5 years of observations)? Such temporal subsets may better represent real-world data-scarce scenarios, where only a limited historical record is available (e.g., in some basins, only the latest 10 years data is available). Evaluating the proposed framework under these conditions could provide additional insights for practitioners developing streamflow prediction models in data-scarce regions.

Response: Thanks for the comments. We performed random sampling of labeled sequences from the Train_{100%} subset, as this strategy has been demonstrated to outperform targeted sampling approaches, e.g., high-flow-only sampling strategy (Heudorfer and Loritz, 2026). Additionally, short-term data gaps may occur in some cases, due to temporary failures of measurement instruments, resulting in periods without streamflow observations within a given year. Such missing-data event can occur irregularly and may be considered relatively random. For example, in basin DE110010 in CAMELS-DE dataset, 8, 50, and 17 daily streamflow values are missing in 1998, 1999, and 2000, respectively. Thus, we adopted a random sampling strategy to simulate data-scarce situations in real-world scenarios.

2. The median NSE values for the in-domain and out-of-domain experiments appear to be very similar, regardless of the amount of labeled data used (Figures 4 and 7). Could the authors provide an explanation for this observation? This result is somewhat surprising, as out-of-domain prediction is generally expected to be more challenging than in-domain prediction due to differences in data feature distributions between the source and target datasets.

Response: We randomly divide all basins from the CAMELS-DE dataset into two groups. The first group includes 1,265 basins used for regional model development and in-domain performance evaluation. The second group includes the remaining 317 basins used to evaluate out-of-domain generalization performance. After reviewing the geometries of the basins, we found that many basins have substantial spatial overlap across a large part of Germany, as shown in the figure below, as shown in the figure below. In this figure, the geometries of the catchments are shown transparently, so a darker colour means that the geometries of the catchments in that area overlap; the darker the colour, the higher the density of catchments in that area.



This spatial overlap may provide an explanation for the similar median NSE values observed in the in-domain and out-of-domain experiments. Specifically, the selection of 317 basins is not strictly based on a spatially separated or climatologically distinct domain, compared to the 1,265 basins. Thus, the source and target basins can still share similar meteorological, topographical, and hydrological characteristics. While the 317 target basins are not seen

during model training, they may still share the similar data features distributions with the 1,265 source basins. We will add this discussion and explanation in the revised manuscript.

3. The authors train models using data across all flow regimes (including both low-flow and high-flow conditions) and evaluate performance using FLV and FHV metrics to assess low-flow and high-flow behavior, respectively. Have the authors considered training SSL or baseline models exclusively on high-flow data and evaluating their performance specifically under high-flow conditions? Such an evaluation could provide additional insight into the models’ ability to learn regime-specific dynamics and their robustness under extreme flow conditions, while it may be outside the scope of the present study.

Response: Thanks for the valuable suggestion. We did not consider to train models only using high-flow data. In streamflow prediction and flood forecasting, it is common practice to train models using data covering all flow regimes (Kratzert et al., 2018; Zhang et al., 2026). Moreover, a recent study (Heudorfer and Loritz, 2026) showed that training LSTM models exclusively on high-flow regimes leads to significantly lower overall performance (in NSE), compared to training models on all flow regimes. Thus, we trained models on all flow regimes, to ensure robust overall streamflow prediction performance. Performance under high- and low-flow conditions was then evaluated separately as a complementary assessment of model performance across different flow regimes.

4. In Line 270, the authors state that the results reported in Sections 4.1 and 4.2 are the median values obtained from 10 runs with different random seeds. However, the hydrographs presented in the manuscript (e.g., Figure 6) appear to show predictions from only a single run. I recommend presenting the ensemble prediction uncertainty across all 10 runs, for example by showing the minimum, median (or mean), and maximum predicted streamflow values, or by including confidence bands around the median prediction. This would provide readers with a clearer assessment of the robustness and stability of the proposed method with respect to random initialization.

Response: Thanks for the comments. We agree with the reviewer that we need to report and discuss the robustness and stability of the proposed method. We will report the metric results across 10 runs using box plots (perhaps in Appendix), and discuss the resulting performance variability. We will also report the performance across 10 runs in the relevant figures.

5. Please use the term flood with caution throughout the manuscript, or clarify that the high-flow events analyzed in this study correspond to actual flood events. If the evaluated events do not necessarily result in flood, terms such as high-flow events or peak flows may be more appropriate.

Response: The events we investigated are “extremes” for these catchments (i.e., high flow or high flow). We agreed with the reviewer that the term “flood” is not strictly correct in this context. We replaced “flood” with “high flow” throughout the revised manuscript.

References

- Heudorfer, B., Loritz, R., 2026. Is smart sampling worth it? impact of training data selection on the performance of lstms in streamflow prediction. *Hydrology Research* , nh2026119.
- 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, 6005–6022.
- Zhang, B., Ouyang, C., Xu, Q., Liu, Y., Wang, Z., Tang, Q., 2026. Temporal inductive biases in hourly flood forecasting: a comparative analysis of recurrent, attention-based, and state-space neural networks. *Journal of Hydrology* , 135727.