

Summary:

From my understanding of the paper, you introduced a semi-supervised learning framework consisting of two steps: an encoder and LSTM are first pre-trained through Contrastive Predictive Coding (CPC), using meteorological forcings and catchments attributes. The model is then fine-tuned with varying amounts of labeled streamflow data to simulate different levels of data scarcity. You compared your approach with a “classical” LSTM architecture, trained with previously defined varying levels of data scarcity. You also experimented with these two methods on single basins. This is a solid contribution in the domain of streamflow prediction in poorly gauged basins, and you also highlight some of the limitations of your approach, which is always appreciated.

However, I see some issues with the current state of your manuscript:

- (1) For the pre-training of your CPC method, you randomly divided your sequences into train and validation. However, since your sequences can overlap up to 11 months, I fear that you could be overfitting without noticing.
- (2) Your fine-tuning training sets rely also on random splitting. As this may not be too much of an issue for low percentages of data scarcity (e.g., $\geq 60\%$ available), this could affect a lot the training and fine-tuning of your models for lower percentages. In return this could false the results and possibly your claims (which I hope not).
- (3) Your training/fine-tuning is performed on a small number of epochs, and your performance curves do not seem to show that convergence is achieved.
- (4) You should be careful with the usage of “flood” in the manuscript or make it clearer that the events you look at indeed led to floods.

I recommend accepting with minor revisions, provided that the following comments are addressed.

Response: We thank the reviewer for thoughtful comments and for taking the time to review our work. We address the five issues raised by the reviewer point by point below.

Major comments:

1. L. 175-182: if I understand correctly, the validation set could contain sequences that overlap by up to 11 months with sequences contained in the train set (since you sample randomly). How did you ensure that you do not overfit during CPC? You should either make it clearer that you addressed this issue already or design a different sampling strategy to ensure that you’re not overfitting.

Response: We thank the reviewer for this concern. Your understanding is correct. Randomly sampling may results in a 11-month overlap between one training sample and one

validation sample for CPC pre-training. As an initial investigation of CPC for streamflow prediction, our work did not specially design the sampling strategy to ensure that the training and validation samples are totally different. To partially mitigate the risk of overfitting during CPC pre-training, we pre-trained CPC models for 100 epochs and selected the model weights corresponding to the minimum validation loss (see Line 210-215, Page 10). Then, we used these selected weights for the downstream fine-tuning task. As shown in Figure. B1 in Appendix, validation loss consistently decreases throughout the pre-training process, and we did not observe evidence of overfitting on the selected validation subset. Despite this constraint, the proposed approach achieves competitive performance, demonstrating the potential effectiveness of the methodology for streamflow prediction in data-scarce regions.

We will discuss this as a limitation of the present study. We will indicate that such sampling strategy is not a standard practice in streamflow prediction tasks. Future work could investigate alternative sampling strategies to construct more strictly independent training and validation subsets for CPC pre-training. That may lead to a more robust CPC model, further benefiting the downstream streamflow prediction task.

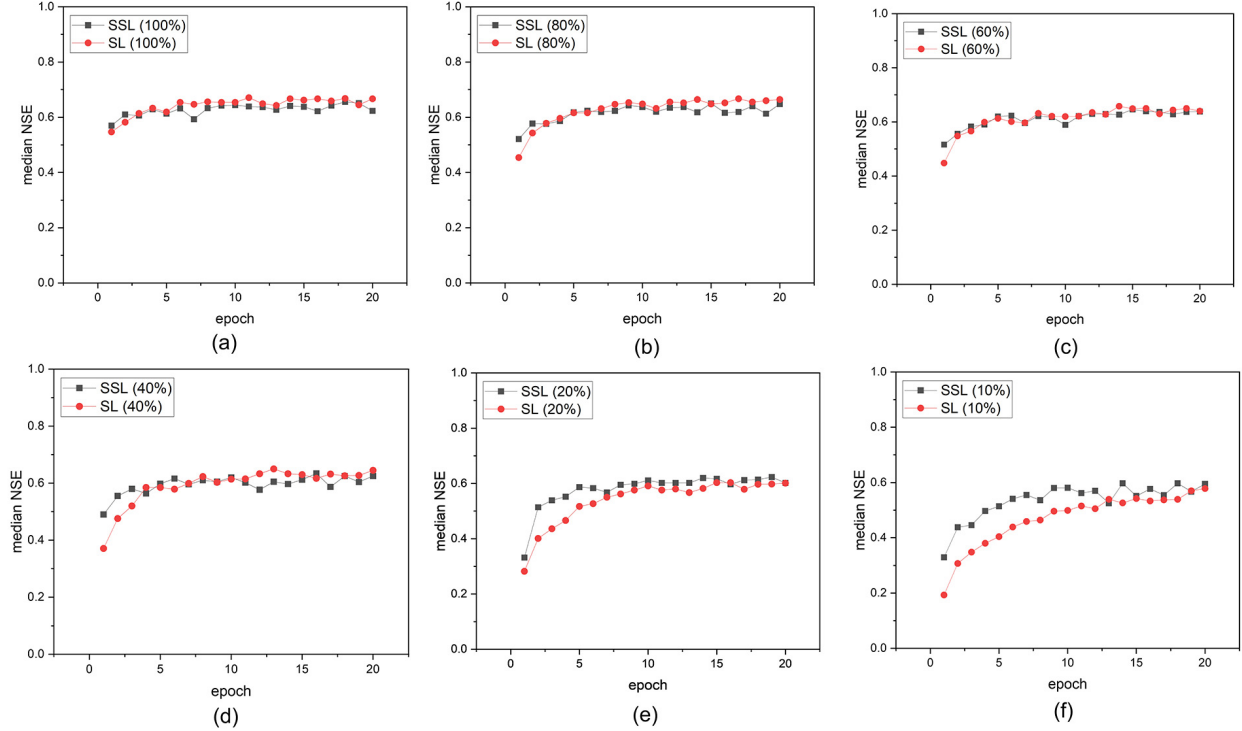
2. L. 187-189 and L. 271: My understanding is that the different seeds are only used for the training/fine-tuning phase. This is not an issue for low levels of scarcity (e.g., $\geq 60\%$ available data), but this could change by a lot the results you get for high levels of scarcity. (1) If the seeds were used as well in the data splitting strategy, you should make it clearer when introducing your different training sets; (2) If not, you need to provide some measure of uncertainty for your performance, as it could change greatly depending on the data present in the split. I would assume that computation-wise, it shouldn't be an issue to run multiple times this training/fine-tuning for higher levels of scarcity.

Response: Thanks for the comments. The seeds are not used to split datasets. The same training, validation and test subsets were used across all runs with different seeds. We reported the median values across 10 runs in Experiments 1 and 2. We agree with the reviewer that we need to report and discuss the performance uncertainty. We will report the results across 10 runs using box plots (perhaps in Appendix), and discuss the resulting performance variability.

3. Figure 5.: Related to my first comment. It is unclear from Figure 5. whether any of the methods reached convergence. It seems to me that in most cases, none of the networks has reached convergence (high variability in some cases, score increasing in others). It also seems that the SSL method does not stabilize for some of the data splits. You should make sure that both networks achieve convergence to have a fair comparison. You should also include the curves for the three highest levels of scarcity (possibly in the appendix) to make sure that you are doing things correctly. A suggestion from my side would be to use a higher number of base epochs (e.g., 50) and to implement an early stopping strategy.

Response: We originally presented this figure with a narrowed y-axis range for median NSE to better highlight the differences in validation performance between SSL and SL during the

early stages of fine-tuning (e.g., 0.54-0.68 for 100% labeled data). This highlights improved training efficiency of SSL, that initializes model parameters with pre-trained weights on the relevant meteorological forcing data. These pre-trained weights provide a more informative starting point than random initialization, facilitating faster convergence and reducing the risk of overfitting. As a result of the narrowed y-axis range, the validation performance in the final epochs may appear to be “unstable”. However, the actual differences in NSE across the final epochs are very small, as demonstrated in our revised figure using the standard NSE range of 0–1 (see below figure). From this revised figure, we can see that the models have generally reached convergence in most cases, suggesting that extending the training for additional epochs is unlikely to provide substantial benefits. Additionally, we used the same number of epochs as well as other settings with those used by [Loritz et al. \(2024\)](#), who developed the benchmark model using the CAMELS-DE dataset. We will use this new figure in the revised manuscript. We will add a figure to show other three highest levels of scarcity (i.e., 5%, 1% and 0.5% labeled data) in the appendix.



4. All relevant figures: most figures showing model performance show only one curve (for seed=110). This choice can raise questions about reproducibility and stability. To convince me, you should either include a “stability” curve (e.g., mean $\pm$ std across seeds) or move all per-seed figures to the appendix (or plot all remaining curves on one figure in the appendix).

Response: We will report the prediction results across the 10 runs in the relevant figures. We may include additional figures in the Appendix and retain the original figures in the main manuscript to facilitate a clear presentation in model performance, while they include the median metric value of 10 runs or prediction results from one seed.

5. General comment: you use streamflow as your indicator for “flood” events. Are the events you investigate related to real floods? I’m not aware of the general conditions and values of streamflow in Germany but I think you should make it more explicit. You should also make it explicit if these catchments are prone to high-streamflow events (e.g., the Rhein falls in Switzerland would make it into this category) or if the events you investigate are “extremes” for these catchments. This relates to my minor comment about Fig 6. and B3.

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.

Minor comments:

1. Table 1. and Table 2.: It was hard for me to follow how the numbers of sequences in the test and validation were obtained. It is also unclear to me why the time period for the third row shows only 1999. From the number of sequences, my understanding is that training/fine-tuning occurred on the full 1970-1999 period. I think the captions of the tables could be improved to provide more information.

Response: In terms of Table 1, we first strictly followed the training (1970-10-01 to 1999-12-31), validation (1965-10-01 to 1970-09-30) and test (2000-01-01 to 2020-12-31) periods used in the development of the benchmark model in [Loritz et al. \(2024\)](#). Then, we divided this training period into two periods, and generated (1) Pre-train_{CPC} (1970-10-01 to 1998-12-31), and (2) Train_{100%} (1998-12-31 to 1999-12-31) subset. They were used for CPC pre-training and fine-tuning, respectively. This split can ensure the Train_{100%} subset (with data from 1999) can initially represent a “data-scarce” region. In Table 3, you can find that two subsets: (1) Pre-train_{CPC} with unlabeled data from 1970-1998, and (2) Train_{100%} with labeled data from 1999. We have also indicated the details of these datasets in the context. Time period division method for Table 2 is similar with that for Table 1. We added the clarification on period division in the revised manuscript.

2. L. 189: (1) How did you choose your levels of data scarcity? (2) Are they representative of any real-world situation? You could make this more explicit. (3) Did you perform any check in the spatio-temporal coverage that these sets of data represent? This relates also to my second major comment.

Response: (1) To the best of our knowledge, we did not identify any previous study in hydrology that explicitly defines different levels of data scarcity based on specific amounts of available training data. Thus, we can only simulate different levels of data scarcity using our current procedure, in the absence of an established definition. (2) We believe that these levels could represent some real-world data-scarce situations in certain regions. For example, [Ouyang et al. \(2025\)](#) developed a single-basin model using 7 years of data available (a maximum of 2,191 sequences) from a basin in China. This is considered as a data-limited scenario by [Ouyang et al. \(2025\)](#). This level of data availability is comparable to our most data-scarce setting (Train_{0.5%} with 2,057 sequences). I added this discussion in

the revised manuscript. (3) We did not perform any check in the spatial-temporal coverage for each small subset. In this study (Experiment 1 and 2), our primary objective was to investigate the performance of a regional model under different levels of data scarcity, with data scarcity represented by the amount of available labeled data. We did not aim to evaluate the performance of individual basins at their respective levels of data scarcity.

3. L. 271-272: my understanding is that you show the median of the performance (i.e., calculate all performances then take the median score). If this is what you mean, I’m fine with the phrasing. If you mean instead the performance of the median (i.e., train across seeds, calculate the ensemble median predictions and evaluate), you should adapt your phrasing to make it clearer.

Response: To evaluate regional model performance, we mainly calculated the median metric values across basins for each run. Then, we obtained 10 median metric values from the 10 independent runs with different random seeds, and reported the median of these 10 values in Sections 4.1 and 4.2. We revised the context to make it clearer.

4. Figure 4. and Figure 7.: I understand you show the median performance. Did you observe substantial differences between your different “seed runs” or are the methods stable across seeds? You could add error bars for the performance range. This also relates to my comments on figures showing performance for only one seed.

Response: We agreed with the reviewer. We will show the performance across different runs and discuss the results.

5. L. 302: given the curves shown in Figure 5., I would be careful with the claim about overfitting. It is unclear to me if your SSL does indeed improve on that regard since the validation score shows more unstable behavior. This relates to some extent to my third major comment. Depending on the “new” results, you may have to adapt (tone down) the claim or totally remove it.

Response: See my response to the third major comment. The revised figure show that both of SSL and SL model training are stable in most cases.

6. L. 303: From what you wrote in L. 210-211, I would assume the pre-training is the computationally expensive step of the method (although this scales well since you then fine-tune the same base architecture). You could maybe add more details (e.g., table in the appendix) about the resources (mostly training time, since hardware is the same) to strengthen your claim.

Response: Thanks for the comments. We added the details to strengthen my claim. We indicated that fine-tuning models on Train_{100%} dataset takes less than 2 minutes per epoch in our work. It is not computationally expensive, compared with CPC pre-training (1 hour per epoch).

7. Figure 6. and Figure B3. (and L. 316-317): You could convert mm/day to m³/s (or any flow units) to strengthen the claim about “highest streamflow events” (L. 316-317).

Response: We think converting unit is unnecessary. The current unit (mm/d) is a common unit used in hydrograph (Kratzert et al., 2024, 2018). The high flow reported in these figures represent one of the top five highest flow events in one basin, regardless of whether their corresponding values are large or small after conversion to m³/s.

8. Figure 6. and Figure B3. (and L. 316-317): I did not understand to which dataset belonged the two basins DE212760 and DE710280. (1) If they are in test_{out}, you could add a sentence about it. (2) If they are in test_{in}, I would be interested in knowing how different from the training distribution these events are. This could explain the (almost) perfect modeling with the 100% training.

Response: These basins are included in Test_{in} subset. A detailed analysis of the data distribution is beyond the scope of this study. This work by Acuña Espinoza et al. (2025) may be interesting for you.

9. Figure 6.: The SL performance increases as the amount of training data increases, which is consistent with previous studies. However, your SSL shows a small drop from 20% to 40% and a bigger drop from 60% to 80% before increasing again. Did you investigate this behavior with larger case studies? Could this be linked to my comments on convergence of the training?

Response: We did not systematically investigate this behavior across a large number of high-flow events and basins. We do not think that this behavior is attributed to the convergence of the training process (see our response to the third major comment). Instead, we think that it may be related to differences in the level and number of high-flow events included in the different training subsets. These subsets were generated through random sampling from a small dataset (Train_{100%} with one-year data). Thus, a high-flow event occurring in the Test_{in} subset, may be represented in Train_{20%} but not in Train_{40%}. Such differences in the high-flow events available for training could contribute to this behavior.

10. Figure B3.: we observe that the performance does not increase too much for the classical approach when increasing training data. Could this relate again to the convergence? It could also relates to the comment above about the dataset in which this basin is.

Response: See my above response. We think this high flow event in Test_{in} subset (see Figure B3) may be higher than those in the Train_{100%} subset. The model may not generalize well to such higher flow, regardless of the amount of data used for training (e.g., Train_{40%} to Train_{100%}), as inspired by Acuña Espinoza et al. (2025). Thus, as the amount of training data increases, the model’s generalization ability to such higher flow does not improve substantially.

11. Table 3., 4. and 5.: add units for the metrics that are not unitless. I may have missed

other places where this is needed.

Response: Thanks for the remind. We added the units for MSE, RMSE, and peak-timing error.

12. Section 4.3: I think the comparison is unfair to the LSTM. We know from previous research that one should not use LSTMs in such context (single basin, 1 year of data) as they don't perform well. You could remedy this in multiple ways: (1) Use a process-based model calibrated on your basins as the baseline; (2) Use more data (e.g., 3-5 years). Although this is still very limited, it could make for a fairer comparison; (3) Drop the section and move it in the discussion of limitations (your method seems to also suffer from the lack of data, see Figure 8.).

Response: While we agree with the reviewer that we should train LSTMs with more data and from more basins, we designed Experiment 3 as a supplementary experiment to Experiment 1. In Experiment 1, we considered a data-scarce scenario with at least 2,057 sequences available ($\text{Train}_{0.5\%}$). However, this setting may not fully represent more extreme data-scarce scenarios that can occur in the real world. Thus, we added a more extreme scenario with a very limited number of streamflow observations (366 sequences), and compared model performance of SSL and SL in such scenarios in Experiment 3. The comparison results could provide insights into which method (SSL or SL) is more effective, when the number of available streamflow observations ranges from 366 to 2,057 sequences. While our method suffers from the lack of data (as shown in Figure. 8), our method significantly outperforms the baseline model. The focus of this experiment was not to develop models for actual deployment, but to show clear evidence that the SSL methods are more effective than SL in more extreme data-scarce scenarios. Furthermore, all settings for SSL and baseline SL are same, except that SSL additionally used the unlabeled data for pre-training. Thus, we think this comparison is fair. We would like to retain this section and Experiment 3, as we consider its experimental design appropriate for the purpose of this study.

13. Table 5.: I would be interested in knowing the number of peaks you evaluate on.

Response: We did not explicitly calculate the exact number of peak-flow events used for evaluation, as this also depends on how peak flows are defined, i.e., the threshold above which a flow value is considered a peak flow. Instead, we used three common metrics: FHV, peak-timing error, and $\text{MAPE}_{\text{peak}}$. For example, FHV reflects the model ability in predicting the top 2% peak flow range in test subset, i.e., around 153 (high) flow values per basin is calculated.

14. L. 390-391.: You could tone down your claims as Figure 9. does not support them. While your method shows "better" time series, it does not capture the trend nor the flow peak.

Response: We agree with the reviewer. We will revise this paragraph to adopt a more impersonal tone.

Notes:

1. L. 167: you describe CAMELS-DE as providing data starting in 1951, but you only use data from 1965 onwards. Is it because the data is not complete for these years or is there another reason (I’m not an expert of this dataset)?

Response: We thank the reviewer for raising this point. Although CAMELS-DE provides data from 1951 onwards, we found that the observed streamflow data are more consistently recorded from 1 October 1970 onwards. Additionally, we followed the training, validation, and test periods adopted by the CAMELS-DE benchmark model developed by [Loritz et al. \(2024\)](#).

2. L. 240: you could drop either RMSE or MSE since they represent the exact same information but scaled. You could maybe replace one with the overall bias of your model.

Response: We agree with the reviewer that MSE and RMSE are mathematically related. However, we would prefer to retain both metrics to facilitate direct comparisons with other studies, as other researchers may report either MSE or RMSE.

3. Section 4.3, Figure 8.: while you improve over the classical approach, the results for the first 5 basins show NSE values around 0.1, which remain quite low. (1) This highlights that SSL struggles too in this context (this can also be seen for the instability in performance depending on the catchment). (2) This relates to my minor comment about the need of this section, which could be discussed less extensively in the discussion.

Response: Thanks for the comments. We will revise this paragraph to adopt a more impersonal tone. We will highlight that both methods significantly underestimate peak flow, while the SSL predictions are more closely aligned with the observations compared to SL. Based on our responses to the minor comment (No.12), we would like to retain this section and the Experiment 3, as we consider its experimental design appropriate for the purpose of this study.

4. L. 422-428: From my understanding, classical transfer learning approaches can transfer a relation between inputs (meteorological forcing, geographical data, catchment attributes) and outputs (observations of streamflow e.g.). On the other hand, SSL transfers only an encoding of meteorological “behavior” and geographical/catchment attributes. You are right that transfer learning should be first trained on a variety of catchment representing the catchment you want to transfer to. However, when searching for relevant papers, I came across a paper from [Alzhanov et al. \(2025\)](#), where they seem to address some of the concerns you mention about transfer learning. This could be an addition to balance your claims. Note that I read rapidly this reference.

Response: Thanks for the reference recommendation. After reviewing, we found they pre-trained models on a global dataset Caravan, and fine-tuning it using the local data. Their results show it enhance the prediction performance ([Alzhanov et al., 2025](#)). However, a recent

study (Gao et al., 2026) showed that transfer learning performance for a specific target may be hindered by irrelevant source inputs and suboptimal fusion of source information. Reducing the domain mismatch between the source and target domains is therefore an important research topic that needs further investigation. Compared with identifying and adopting an appropriate transfer learning strategy, our SSL method provides an alternative approach to improving prediction performance by leveraging relevant unlabeled data from the target basin.

5. L. 525-526: You used the base values as those provided in the code of NeuralHydrology. Have you tried different approaches, e.g., by reducing the number of timesteps between peaks and increasing the prominence? With this, you could study how your method performs in case of multiple peaks in a short timeframe. I know you would have to go into the NeuralHydrology code, so I leave this as a note.

Response: No, we did not investigate other approaches or evaluation metrics in this study. They could be considered and investigated in future work.

6. Figure B3.: some of your labels are partially hidden (does not affect understanding)

Response: Thanks for the remind. We revised the figure to show labels clearly.

References

- Acuña Espinoza, E., Loritz, R., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., Ehret, U., 2025. Analyzing the generalization capabilities of a hybrid hydrological model for extrapolation to extreme events. *Hydrology and Earth System Sciences* 29, 1277–1294.
- Alzhanov, A., Nugumanova, A., Moreido, V., 2025. Transfer learning using the global caravan dataset for developing a local river streamflow prediction model. *Environmental Modelling & Software* , 106691.
- Gao, Y., Liang, D., Borgomeo, E., Cho, H., 2026. Multi-source adaptive-fusion transfer learning for streamflow forecasting in data-scarce catchments. *Water Resources Research* 62, e2025WR042495.
- Kratzert, F., Gauch, M., Klotz, D., Nearing, G., 2024. Hess opinions: Never train a long short-term memory (lstm) network on a single basin. *Hydrology and Earth System Sciences* 28, 4187–4201.
- 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.
- Loritz, R., Dolich, A., Acuña Espinoza, E., Ebeling, P., Guse, B., Götte, J., Hassler, S.K., Hauffe, C., Heidbüchel, I., Kiesel, J., Mälicke, M., Müller-Thomy, H., Stölzle, M., Tarasova, L., 2024. Camels-de: hydro-meteorological time series and attributes for 1582 catchments in germany. *Earth System Science Data* 16, 5625–5642.
- Ouyang, W., Zhang, C., Ye, L., Zhang, H., Meng, Z., Chu, J., 2025. Dive into transfer-learning for daily rainfall-runoff modeling in data-limited basins. *Journal of Hydrology* 657, 133063.