< REPLY TO REVIEWER 2>

- **Title: Self-training approach to improve the predictability of data-driven rainfall-runoff model in hydrological data-sparse regions**
- **Authors: Sunghyun Yoon and Kuk-Hyun Ahn**

((Acknowledgement)) The authors sincerely thank the reviewer for their helpful and constructive comments.

((Comment #1))

It is not clear to me why this method is called semi-supervised learning. Fundamentally, the source model is supervised learning and the new model is also trained using supervised learning. While some similar work in AI who used a student-teacher paradigm did call themselves as such, these studies typically used much more unlabeled data and the structures within. Semi-supervised learning typically involves leveraging a small amount of labeled data alongside a larger pool of unlabeled data to improve learning. In this setup, the pseudo labels generated for the new site's data act as a mechanism to utilize unlabeled data. However, in the case of this paper, it is not clear to me what information is leveraged from the target domain's data, especially the other unlabeled data points.

((Reply))

According to the literature in the field of machine learning (e.g., Chen et al., 2022; Learning, 2006), we can define semi-supervised learning as a framework that leverages both labeled and unlabeled samples to augment the available training data for models. Additionally, it's worth noting that reviewer #1 shares a similar interpretation (refer to comment #1 from reviewer #1) and highlights that our approach is fundamentally rooted in self-training, a subdomain of semi-supervised learning. While we could consider the concept of a student-teacher model for our proposed method, it might not sufficiently emphasize our unique approach that utilizes unlabeled data and introduces a novel loss function designed to handle two distinct types of samples. In practice, whether a given learning paradigm is supervised or semi-supervised does not depend on the use of teacher-student framework. Teacher-student framework can be used with both supervised and semi-supervised learning methods, which depend on the task conditions. Nevertheless, we acknowledge the reviewer's concerns, and we have addressed this information in the revised manuscript.

Revision in the manuscript (Lines 229-231):

"While our framework exhibits characteristics of a student-teacher paradigm, its innovative use of unlabeled data and the accompanying procedures correspond to a self-training-based framework."

((Comment #2))

The biggest technical problem as I see is that I cannot independently verify that their baseline performance is state-of-the-art. They also lacked comparable results to any other studies. Hence I cannot tell you if the benefits are truly as claimed. Since their primary target of comparison is transfer learning (Ma et al., 2021), they should try to compare with that paper. If they cannot, another paper seems to work on CAMELS and may be comparable is this one (Feng et al. 2021, doi: 10.1029/2021GL092999).

((Reply))

The reviewer's observation regarding the necessity of verifying the baseline performance is accurate. In accordance with the agreement, we made an effort to replicate the models used in the transfer learning study conducted by Ma et al. (2021). Following the training and testing scheme outlined by Ma et al. (2021), our study involved two training scenarios: a 1-year and a 5-year training scenario. To be specific, in the 1-year (5-year) training scenario, the models were trained from January 1, 2004, to January 1, 2005 (January 1, 2000, to January 1, 2005), and subsequently tested from January 1, 2005, to January 1, 2010 (January 1, 2005, to January 1, 2010). Our experiments yielded performance statistics for the test phase in our study (refer to Table S2 in our revised manuscript), which bear resemblance to the results reported in Ma et al. (2021) (specifically, their Table S3). While our basin selection, which comprised 666 basins for the 1-year scenario and 668 basins for the 5-year scenario, slightly differs from that of Ma et al. (2021), who employed 667 basins in both scenarios, we are confident that our study has acceptably validated the baseline performance of the transfer learning approach. More details can be found in text S2 in the supplementary material.

((Comment #3))

There should be at least two alternative approaches: (i) directly training using all the training data and just make forward runs on the test basins; (ii) transfer learning. It is not clear to me if their named experiments (rgn-LSTM) use either of these two, and is just their student-teacher approach.

((Reply))

This comment primarily pertains to the third experiment among the three conducted in our study. We acknowledge the importance of comparing our proposed approaches against two alternative approaches, which are: (i) direct training using all the available training data and performing forward runs on the test basins, and (ii) employing transfer learning. The results of the comparative analysis between our proposed approach and the first alternative approach are presented in the first experiment, considering two model settings: the individual (ind-) and regional (rgn-) settings. You can find these results in Figures 4 and 5. Moreover, the comparison between our approach and the second alternative approach (i.e., transfer learning) is included in the third experiment. Specifically, the third experiment is designed to assess the performance of our proposed framework in comparison to separate training approaches. We have also incorporated two pre-training models, followed by fine-tunings, in this experiment. To summarize, we concur with the reviewer's recommendation that our proposed approaches should be compared with the two alternative approaches suggested. These comparisons have been presented in the results section of our study.

When it comes to their nomenclature (i.e., rgn-LSTM-xx), the initial two components are tailored to the specifics of our experimental context. To illustrate, the first element (i.e., rgn) conveys information regarding the regional model configuration, whereas the second element (i.e., LSTM) denotes our model's architecture. It's worth noting that all the alternative approaches mentioned by the reviewer are trained based on these experimental setting.

((Comment #4))

It should be acknowledged that transfer learning can use different input items across source and target regions, even different amounts of inputs. The authors' approach cannot allow this (without additional changes).

((Reply))

We agree. The approach of transfer learning can yield benefits when employing distinct input data sets for source and target regions, even when varying amounts of inputs are necessary. This information has been added in the revised manuscript.

Revision in the manuscript (Lines 536-538):

"For example, transfer learning can yield benefits when employing distinct input data sets for

source and target regions, even when varying amounts of inputs are necessary."

((Comment #5))

The experimental design and comparisons are very confusing and difficult to remember. There are so many different versions, experiments and acronyms and it was a torture to demand readers or reviewers to remember all these. You should more clearly present the core comparison, making it really easy to see the benefits, and then expand on more experiments. You can also consider removing some unimportant experiments.

((Reply))

The point is well taken. The outlines of the manuscript in sections 1 and 2 have been updated for clear explanation. As we declared in the introduction section, this study specifically explores the following hypotheses:

1. **The effect of self-training (in Sections 2.4.1 and 3.1)**: The availability of additional climate data, i.e. unlabeled data, could potentially enhance the performance of LSTM models in producing reliable streamflow predictions in diverse modeling scenarios.
2. **The effect of annealing process (in Sections 2.4.2 and 3.2)**: It would be beneficial to use a self-training-based framework that leverages both labeled and unlabeled data, but treats them differently instead of treating them homogeneously. By differentiating the weights of labeled and unlabeled data and incorporating them into the training process, the model can potentially achieve better performance on unseen data.
3. **The effect of joint training and avoiding domain mismatch (in Sections 2.4.3 and 3.3)**: The joint training of both labeled and unlabeled dataset enables the loss terms for labeled and unlabeled samples to be jointly optimized. It has the potential to improve model performance in comparison to a separate training approach (i.e., pre-training followed by fine-tuning) prone to overly bias toward the distribution of samples used in either the pre-training or fine-tuning step.

To address these three hypotheses, the massive experiments are required, resulting in the utilization of numerous acronyms within the manuscript. We firmly believe in the significance of each experiment and its merit for inclusion in the report. Nonetheless, we remain open to refining our experimental design. If required, we welcome specific guidance from the reviewer on which experiments could be excluded. To further address the concern, it is imperative that the revised manuscript provides a more comprehensive representation of our experimental design. In alignment with this objective, we have incorporated an additional figure into the revised manuscript, which is presented as Figure 3.

New figure in the revised manuscript:

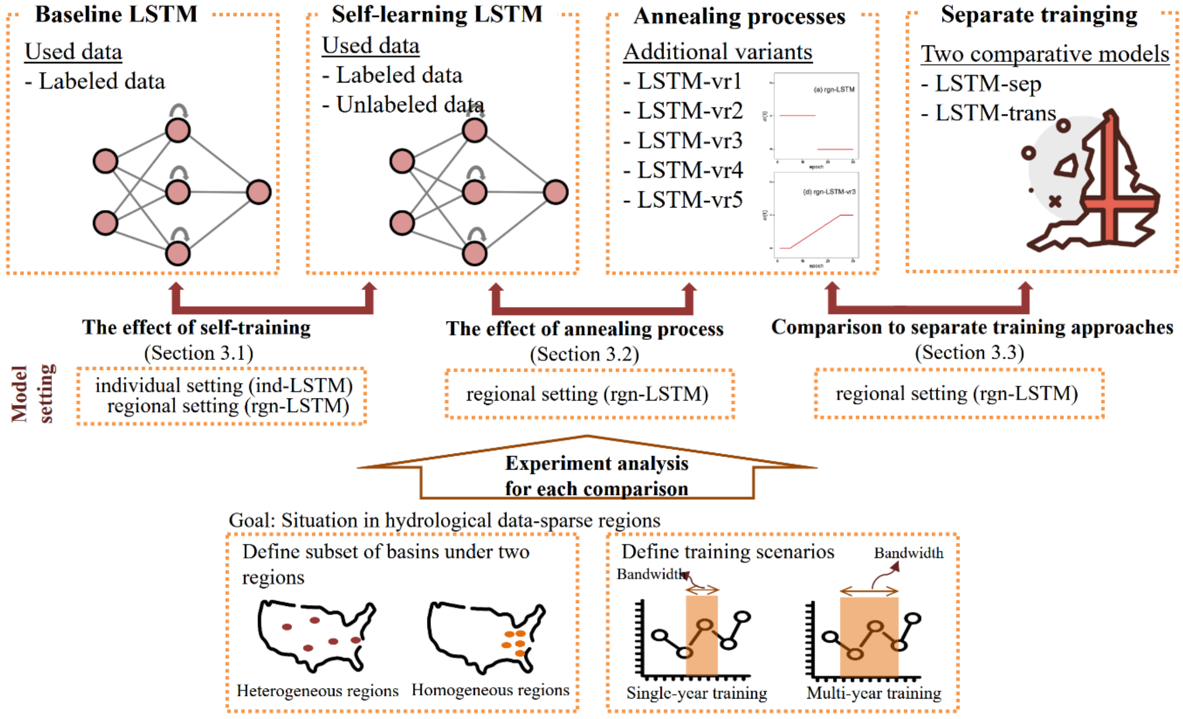

Figure 3 The overview of experimental design.

((Comment #6))

When you apply student-teacher paradigm, the source dataset should be a large and diverse data. UK does not have a very diverse geography. It would make more sense to use CAMELS USA as the source data. You may see different comparisons in that way.

((Reply))

The potential variation due to diversity has been addressed in the revised manuscript.

Revision in the manuscript (Lines 528-531):

"Additionally, it's worth noting that the performance of transfer learning may be influenced by potential variations such as the selection of the number of regional basins, which reflects the diversity in source data during the pretraining process."

((Comment #7))

The organization of the paper is very poor. You have to dig in carefully and read all the way to Section 2.3 to get the main idea "... the pseudo labels for unlabeled dataset are generated from a pre-trained teacher model trained on labeled dataset. Student model is trained in supervised manner on both the labeled and (pseudo label assigned) unlabeled datasets." This should be clear in the abstract and related work should be mentioned in the abstract.

((Reply))

Thank you for this comment. In response to this comment, the information suggested by the reviewer has been incorporated into the abstract.

Revision in the manuscript (Lines 32-39):

"To fill this gap, we present self-training, a semi-supervised learning approach that imputes the pseudo streamflows for unpaired (i.e., unlabeled) samples to increase the amount of available paired samples. To elaborate, we adopt teacher-student framework. The teacher model is first trained on (limited number of) paired samples and then works as a generator of pseudo streamflow for unpaired samples. The student model is trained on both paired and pseudo streamflow-endowed samples. Notably, our framework introduces an annealing-able loss function for training the student model, designed to compensate for the uncertainty in pseudo streamflow."

((Minor comment #1))

Earlier papers (Gauch et al., 2021 as cited, and Fang et al., 2022, doi: 10.1029/2021WR029583) have already examined how to form the training dataset. The general conclusion is that one should use all the training dataset, and the more diverse and large the training data, the better. hence some sentences for the motivation mentioned in the paper need to be revised.

((Reply))

To respond to this comment, we have modified the sentence in the introduction to emphasize that LSTM performance can be significantly enhanced when ample and diverse training data is accessible. Additionally, we have included the references recommended by the reviewer in the revised manuscript.

Revision in the manuscript (Lines 86-87):

"Notably, several studies have demonstrated exceptional LSTM performance, especially in

situations where abundant and diverse training data are available."

((Minor comment #2))

Their introduction should clearly direct the readers to understand why the unlabeled data are useful.

((Reply))

Thank you for the suggestion. We have incorporated the additional information to understand why the unlabeled data are useful.

Revision in the manuscript (Lines 112-114):

"Nonetheless, valuable insights can be gained by incorporating the remaining climate data into the training process to enhance model performance through supplementary training."

((Reference))

Chen, Y., Mancini, M., Zhu, X., & Akata, Z. (2022). Semi-supervised and unsupervised deep visual learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Learning, S.-S. (2006). Semi-Supervised Learning. *CSZ2006. Html*.

Ma, K., Feng, D., Lawson, K., Tsai, W.-P., Liang, C., Huang, X., et al. (2021). Transferring hydrologic data across continents–leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resources Research*, *57*(5), e2020WR028600.