< REPLY TO REVIEWER 1>

- **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))

The title of the article reads "Semi-supervised learning approach", yet the core of the methods is fundamentally rooted in self-training. This process involves generating pseudo-labels with the model, which are then utilized as new training data to further optimize the model. A revision to offer a more precise description is recommended.

((Reply))

The reviewer's observation about the fundamental basis of our approach being rooted in self-training is correct, and we concur with the recommendation. Consequently, we have revised the title to “Self-training approach to improve the predictability of data-driven rainfall-runoff model in hydrological data-sparse regions”. Likewise, we have made adjustments to certain sentences, such as those found in lines 33 and 36.

Revision in the manuscript (Lines 34-38):

“The self-training approach, which is an emerging machine learning paradigm that additionally incorporates unpaired samples, has the potential to be a highly effective method for modeling rainfall-runoff relationships. In this study, we present a novel self-training-based framework for rainfall-runoff modeling.”

((Comment #2))

The narrative in the methods section is somewhat disorganized. It would be beneficial to restructure the sections pertaining to the model and experimental design for greater clarity.

((Reply))

In reference to this comment, we are somewhat uncertain about the reviewer's intention. Nevertheless, taking into account other comments, we have three solutions to enhance the method section. First, we have inserted a guide at the beginning to introduce the specific

experimental designs. Additionally, we have included a visual representation of the experimental designs to enhance clarity. Lastly, additional explanations have been included for certain sentences in order to deliver information more clearly (e.g., lines 276 and 299). Overall, those modifications can be found in lines 269-307.

In addition, we are receptive to making adjustments to the narrative in the methods section. However, we believe that the structure of the section is acceptably organized. If possible, it would be helpful if the reviewer could provide a more specific description of the section should be re-organized so that we could try to improve the section in a targeted way.

Revision in the manuscript (Lines 269-273):

"This study employs a series of three sequential experiments to investigate our research hypothesis in data-sparse regions. Figure 3 offers a conceptual overview of our experimental designs, while the specifics of each experiment are elaborated upon in the following subsections. For each research hypothesis, this study considers two dimensions to depict the situation in hydrological data-sparse regions."

Revision in the manuscript (Lines 275-276):

"The approach is adopted since, in data-scarce regions, the numbers of the streamflow gauge are also limited in reality (i.e., the spatial-domain diversity is limited)."

((Comment #3))

Regarding lines 223-226: Could you elaborate on what is meant by the "pre-trained model" and how it was sourced? This aspect may have implications for subsequent performances of idv-LSTM and rgn-LSTM.

((Reply))

In response to this comment, we have recognized that the term "pre-trained model" is redundant and may cause confusion for readers. It's important to note that the teacher model is not pretrained within the framework. Consequently, we have omitted this phrase in the revised manuscript.

((Comment #4))

Lines 285-287: Please provide a brief rationale for the choice "Ψ takes on values of 10, 30, and

50".

((Reply))

Additional information regarding the rationale for this choice has been inserted. It can be seen in lines 292-295.

Revision in the manuscript (Lines 292-295):

"While these values have been arbitrarily chosen, they are drawn from previous research, such as Lee and Ahn (2022) and Leisher et al. (2016), which used approximately 30 stations to represent a regional basin network."

((Comment #5))

Lines 289-298: The current explanation seems somewhat ambiguous. It might be helpful to provide a visual representation of the framework and further clarify the model's approach to handling both labeled and unlabeled data during its training and validation phases, inclusive of the pre-trained model.

((Reply))

The point is well taken. Based on this comment, we have incorporated an additional figure into the revised manuscript to better illustrate the experimental design overview. You can now find this figure in Figure 3 in the revised manuscript.

Revised figure in the manuscript:

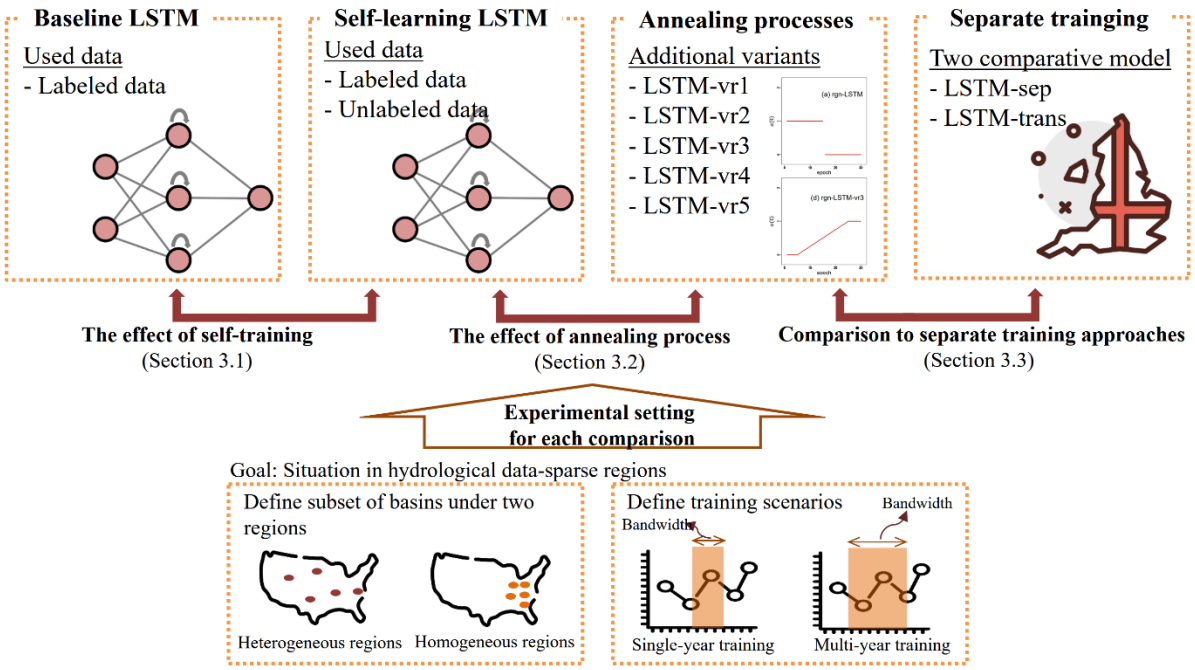

Figure 3 The overview of experimental design.

((Comment #6))

Lines 318-321: While the conclusion mentions a baseline model, but the related information is insufficient, and it's hard to judge the resulting evaluation.

((Reply))

To respond to this comment, we have included additional supportive sentences to introduce a baseline model, which can be found in lines 329-332.

Revision in the manuscript (Lines 329-332):

"This approach is comparable to the methodologies frequently employed in previous studies (e.g., Boulmaiz et al. (2020)) that utilize LSTM for analyzing the rainfall-runoff relationship. Subsequently, these results are then used as the baseline for evaluating the performance of our proposed self-training-based framework"

((Comment #7))

Lines 358-368: It mentioned the use of CAMELS-GB as a source model. It's essential to note that a transferred source model tends to have more variety than local data. It's feasible when the number of regional basins is between 30-150, but the significance might wane when

considering 600 or more basins.

((Reply))

We agree, and we also believe that exploring it would be of great interest. However, the topic may be out of the research scope for the current study. Thus, we have included a note in the revised manuscript to acknowledge the potential for improvement (lines 526-529).

Revision in the manuscript (Lines 526-529):

"*Additionally, it's worth noting that the performance of transfer learning could potentially be influenced by the selection of the number of regional basins during the pretraining process. However, it is not explored in the current study, as determining the optimal number falls outside the scope of our research.*"

((Comment #8))

As for Figure 7 and its associated descriptions: The enhancements seen in the regional model exhibit fluctuations, with the median of improvements oscillating between 0.010-0.027. Considering the dataset's size and spatial distribution differences, one could question the framework's generalizability, especially when the only difference between a single and multiple years in the teacher model is a mere 1 and 3 years.

((Reply))

You are correct that the enhancements seen in the regional model exhibit fluctuations. Nevertheless, we would like to emphasize the clear improvement in the individual setting as a primary outcome. In the regional setting, it is important to note that the LSTM model learns a broader range of rainfall-runoff patterns from a diverse set of basins, which can somewhat diminish the effectiveness of the self-training approach. Even if each site in the regional setting has data spanning three years, the overall model encompasses data spanning a period of 90 years data. We believe that this multi-year span is sufficiently extensive. The median of our results is still positive for the regional setting, indicating that the models trained by our framework continue to exhibit effectiveness beyond what has already been achieved by regional models trained on diverse basin data.

((Comment #9))

In the results section, the bulk of evaluations and consequent findings hinge on the differential values in the NSE metric to discern between models. Such assessments might offer a skewed

perspective. For instance, a surge of 0.1 embodies different ramifications when transitioning from 0.3 versus 0.8. If there's a lack of thorough evaluation in basin selection (considering spatial variations) and basin count, it insinuates potential constraints in the framework's enhancements.

((Reply))

We employ the differential values in the NSE metric to discern between models (baseline versus our model). It's important to acknowledge that such assessments may introduce potential limitations, as highlighted by the reviewer. Nevertheless, the comparative visualization serves as a valuable means to directly assess the performance between our proposed framework and the baseline model. This approach has also been utilized in previous studies (e.g., Lees et al., 2021), although we acknowledge the concerns raised. In response to these concerns, we have included additional figures, namely Figures 4, S2, and S3, which offer a succinct overview of the overall LSTM performance. Please note that, for the sake of brevity, the spatial distribution of each metric for the rgn-LSTM is not presented in this study. This information can be found in lines 403-408 in the revised manuscript.

Revision in the manuscript (Lines 403-408):

"In this section, we assess the effectiveness of the proposed self-training-based framework in enhancing streamflow predictions. Figures 4, S2, and S3 illustrate the spatial distribution of the each metric (NSE, MNSE, and LNSE) and their differences when comparing idv-LSTM to the baseline models during the evaluation period. Also, Figures 5, S4, and S5 show the spatial distribution of the metric differences for rgn-LSTM. We note that, for brevity, the spatial distribution of each metric is not presented for rgn-LSTM."

((Comment #10))

The figure titles, including Figure 5, seem devoid of essential descriptive content. For instance, the number of basins in the dataset is often left unspecified.

((Reply))

We provided the information, "*Therefore, for the remaining analysis, we will adopt rgn-LSTM particularly with the moderate density network.*" (lines 454-455). However, we agree that the description of each figure should be specific. Accordingly, we have changed the titles of Figures 6 and S6 in the revised manuscript.

((Comment #11))

Figure 2 could benefit from a more detailed depiction of the framework.

((Reply))

Thank you for your suggestion. In response, we have completely changed the figure to incorporate more details of our framework.

Revised figure in the manuscript:

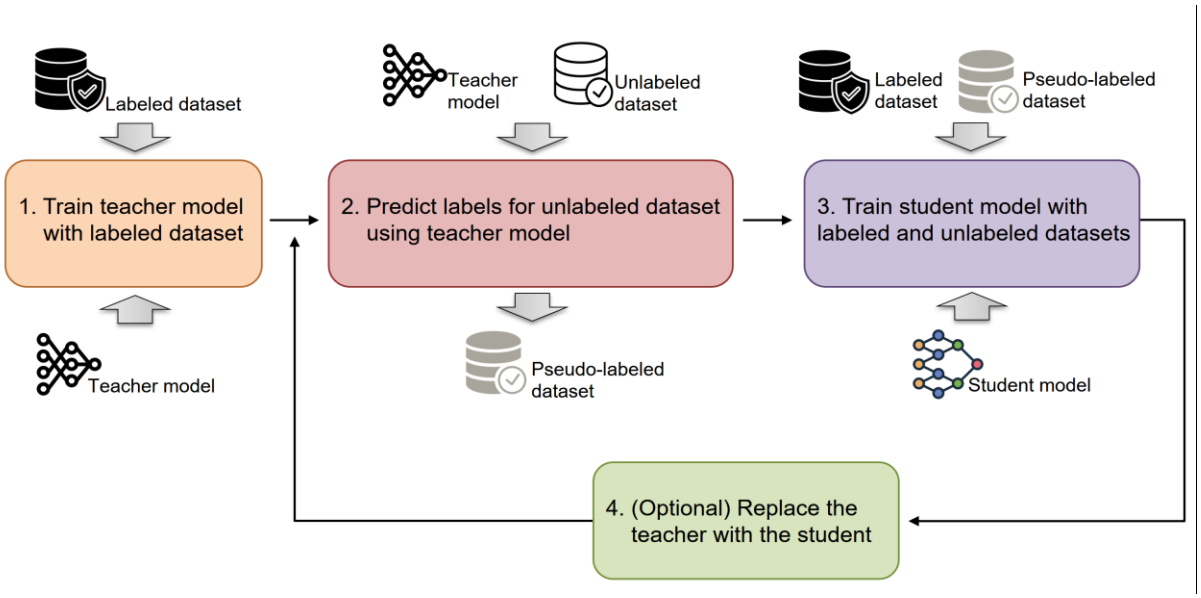

Figure 2 Conceptual illustration of the self-training-based framework proposed in this study. The algorithmic procedure, which corresponds to the numbering described in the manuscript, is also presented.