Reply to Comments

JOURNAL TITLE: **Hydrology and Earth System Sciences**
MANUSCRIPT TITLE: **Incremental learning for rainfall-runoff simulation on deep recurrent neural networks**
MANUSCRIPT-NUMBER: **hess-2024-56**

Dear Editors and Reviewers:

Thank you for your letter and for the reviewers' comments concerning our manuscript. Those comments are all valuable and very helpful for revising and improving our paper, as well as the important guiding significance to our researches. We have studied comments carefully and have made corresponding corrections which we hope meet with approval. Revised portions are highlighted in the paper using the Microsoft Word's "track changes" function. A "clean" version that has accepted all the changes in "track changes" is also provided. A summary of the major changes and item-by-item response to the reviewers' comments are as flowing:

Summary of the major changes:

1. In response to Reviewer 1 and Reviewer 2, we have reviewed the entire article and corrected grammatical errors and rewritten unclear sentences, especially in ***Section 1 Introduction***.
2. In response to Reviewer 1, we have re-explained the significance and innovation of this research method and revised the principle part of the method.
3. In response to Reviewer 2, the first paragraph of ***Section 4.1 Experiment setup*** has been revised to show in detail the setup of the prediction experiments and contrast cases among different models.
4. In response to Reviewer 1 and Reviewer 2, we modified the presentation of the results, converted the tables into figures, and added additional descriptions of the conclusions.
5. In response to Reviewer 1 and Reviewer 2, we have adjusted some of the paper's structures and added relevant information to the methods and experiments sections.
6. In response to Reviewer 1 and Reviewer 2, we re-draw the concept diagram and result diagram of the method for better display and understanding.

Response to Reviewer 2:

1. **Comment:** ***The explanation of the proposed incremental learning method lacks clarity, and the novelty is arguable due to insufficient referencing of prior and similar work. Specifically, the proposed method seems to innovate in how incoming data is selected, yet the lack of references makes it difficult to judge. The explanation provided in section 2.2 is not made sufficiently clear and no reference is provided from line 140 to 215. Moreover, the distinctions between the three proposed incremental learning scenarios are not well defined.***

   Response: Thank you for your insightful and helpful comments on our work. We have added more clear explanation of the proposed method and illustrate the incremental learning scenarios

in more details.

The related part is illustrated in Section 2.2 The Incremental Learning Method.

Existing incremental learning methods for time series data, such as rainfall-runoff data, often do not effectively leverage the temporal characteristics of the data. In this work, we combine data distribution estimation, temporal similarity, and regularization methods to improve the performance of incremental learning for this type of temporal feature data. We utilize partial representative data for incremental training, with a focus on time series similarity metrics that compare time series with the same length. Given the periodic characteristics of rainfall-runoff series, we divide the complete time series into sub-time series of the same length, enabling the similarity between time series with different lengths to be transferred to the similarity among sub-time series with the same length. We ensure that the data in each sub-dataset is similar in distribution and can be fit with a simple distribution, which can be estimated. We integrate similarity in both distribution and time series characteristics as partial representative data selection standards to ensure the representativeness of the selected data. As an additional penalty item on the loss function, parameter importance calculation is the core of regularization during incremental training.

Our method is based on regular network training, and as a result, the amount of calculation is significantly reduced, resulting in a notable acceleration of the training process. Moreover, owing to the representative partial data and regularization, the network model shows good performance on the incremental data. The structure of the incremental learning method can be elaborated as Figure 2, which consists of two main components: regular training for parameter initialization and incremental operation to handle incremental training. Comprehensive consideration of the features of both the historical and incremental data is used to produce partial representative data, reducing the magnitude of the input data. Parameter importance calculation as a regularization constraint is added in incremental training to handle the error problem of the network when training real-time incremental data. Meanwhile, when new incremental data is continuously input, the model may be trained multiple times in a short period. The incremental learning method should also ensure the stability of the method under such conditions.

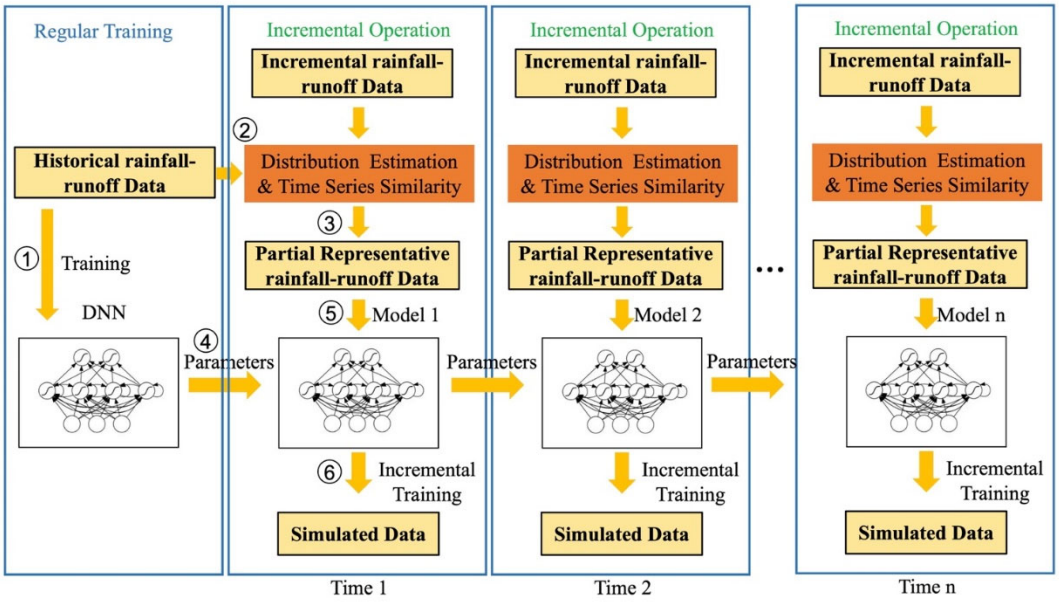

Figure 2: The structure of the incremental learning method.

The incremental operation part can be described in more detail as follows. First, a periodic analysis of the time series is performed, and the combination of historical data and incremental data is sliced into multiple sub-time series. The distribution parameter calculation and temporal similarity measurement are performed for each sub-time series. By comparing the parameter difference between the sub-time series and the overall time series, as well as the temporal similarity difference between the sub-time series, weights are assigned to the calculation results of these differences. This produces the replay scores for each sub-time series. The sub-time series are then sorted according to the replay scores, and the number of sub-time series is determined based on the replay sample size level required for efficient incremental learning. Partial representative samples are selected for incremental training, using the regularly trained network with the parameters initialized. Additionally, during the incremental training process, parameter importance calculation is selected as a regularization constraint and imposed on the training loss of the model. Specifically, L2 regularization is introduced to impose penalties on the loss function of the deep learning model, and the relevant parameters are adjusted accordingly. Finally, the training results are obtained. The process of the incremental operation part is shown in Figure 3.

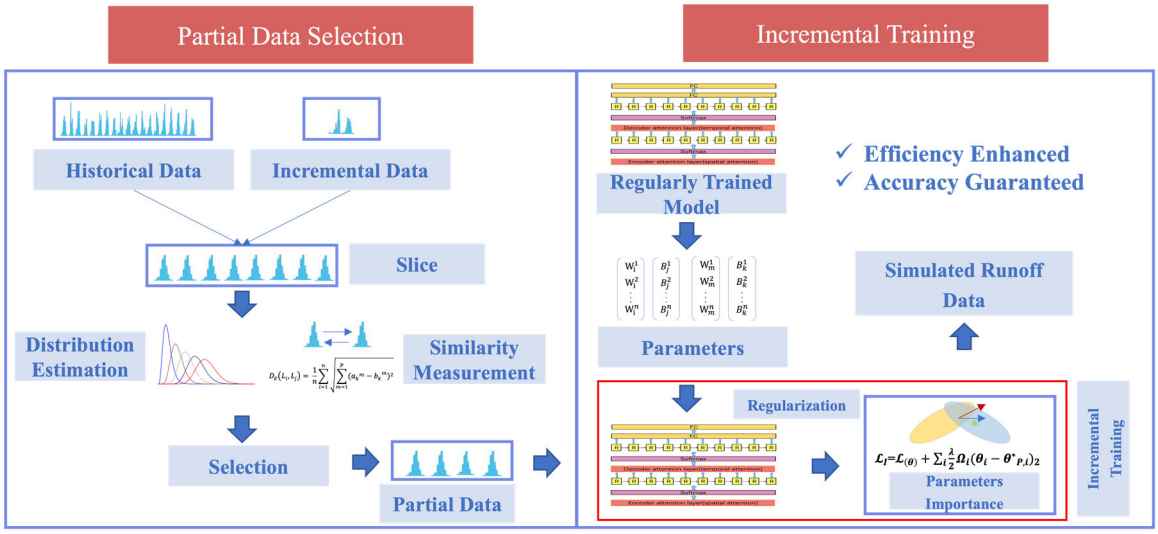

Figure 3: The process of the incremental operation part of the incremental learning method.

Formally, consider the moment t_i, the data that has been processed and trained in the deep learning model is called historical data, denoted as H_{t_i}, the incremental data arriving at this time is denoted as A_{t_i}, the deep learning model is denoted as $M_{t_i}(p_{t_i}^1, p_{t_i}^2 p_{t_i}^3, \dots)$, $p_{t_i}^j$ is the jth parameter of the model at the moment. The historical data and incremental data become the historical data of the moment, and the depth model parameters after training at T_i become the input parameters of the moment model. The complete time series before periodic inspection is denoted as T_W and the sliced time series are T_{t_i}. Skewness and Kurtosis are selected as the distribution estimation metrics and standardized Euclidean distance works as a measure of temporal similarity, the calculation process can be formulated as the following. $Skew$ is Skewness of T_{t_i}, $Kurt$ represents Kurtosis of T_{t_i}.SD is the standard deviation and $\bar{x}$ means the average of T_{t_i}.

$$Skew(X) = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{x})^3}{SD^3} \tag{1}$$

$$Kurt(X) = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{x})^4}{SD^4} - 3 \tag{2}$$

D_E refers to standardized Euclidean distance. α_s, α_k and α_D are the weights correspondent to the metrics to calculate replay score (S_{replay}). Replay score determines the probability that the sub-dataset will be chosen. Skewness and Kurtosis are to measure the distribution difference between $T_{t_i}$and T_W, standardized Euclidean distance is to measure the time series similarity among T_{t_i}. The replay score measures the representativeness of each sub-dataset. Generally, the magnitude of training data is positively correlative to training speed with the same parameters, therefore the incremental learning method can adjust the amount of representative data according to the anticipant speed that the incremental learning method needs to achieve. The number of selected sub-dataset is N_{replay} , and finally such many orders of magnitude sub-datasets with the highest replay scores are selected.

$$D_E(L_i, L_j) = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\sum_{m=1}^{p}({a_k}^m - {b_k}^m)^2} \tag{3}$$

$$S_{replay} = \frac{\alpha_s}{\Delta Skew(X)} + \frac{\alpha_k}{\Delta Kurt(X)} + \frac{\alpha_D}{D_E}(L_i, L_j) \tag{4}$$

Then calculating the importance for each parameter in the network is attached to the loss function of the network, as regularization constraint. This can be described as the following formulations.

$$\mathcal{L}_I = \mathcal{L}_{(\theta)} + \sum_i \frac{\lambda}{2}\Omega_{ij}\left(\theta_i - {\theta^*}_{P,i}\right)_2 \tag{5}$$

$$\Omega_{ij} = \left\|\frac{\partial {l_2}^2(M(x;\theta))}{\partial \theta_i}\right\| \tag{6}$$

$\mathcal{L}_I$ is the loss function of the model during incremental training, $\mathcal{L}_{(\theta)}$ is the loss function of the model during regular historical data training, $\sum_i \frac{\lambda}{2}\Omega_i\left(\theta_i - {\theta^*}_{P,i}\right)_2$ is the constraint item, θ_i is the parameter of incremental meta sample, ${\theta^*}_{P,i}$ is the standard to evaluate the parameter, which represents the difference between the previous and incremental meta sample, Ω_i is the l_2 regularization item, $M(x;\theta)$ is the output of the network, $\frac{\partial(M(x;\theta))}{\partial \theta_i}$ describes the gradient of the loss function of model with respect to parameter θ_i evaluated at the data point x . The importance of parameters can be described as the magnitude of the gradient. And λ can be adjusted with incremental data come.

Uniformly data of early years are set as historical data and data of lately years as incremental data. When the incremental data come at some time, both baseline and the incremental learning method are performed. The rainfall-runoff simulation on incremental data is defined as incremental tasks because the distribution of dataset has changed and three incremental tasks is set on each station. After selecting partial representative data, the incremental learning method uses the regularly trained attention-RNNs with learned parameters and finetuned model with a relatively lower learning rate and part of the changed hyperparameters. As for some of the model parameters, the

weights and biases of the layers are updated when training by back-propagation approach. Iterations are performed with subsets of the training dataset which are called batches or a mini-batches.

2. **Comment:** ***The study's regional focus limits its broader applicability. A model based on studies from a larger area might offer better generalization capabilities and could serve as a more robust baseline to refine.***

   Response: Thank you so much for your insightful and guiding comments. The reason why we chose a specific river basin area for our work is that we hope that our proposed method can effectively solve the key problems in the field of hydrology. In practice, the data patterns of different hydrological research areas lack uniformity, and it is difficult for related deep learning methods to guarantee absolute wide applicability. It would be great if they can be applied to some specific areas.

3. **Comment:** ***The paper deviates from standard dataset division into training, validation, and test splits, common in deep learning, which allows hyperparameter selection and model generalization. Hyper-parameter selection is not done at all and most of the hyperparameters are not reported (e.g. the length of training time series, number of LSTM memory cells, training epochs etc.). The dataset's parameters are vague: where and when is it trained on? The reported results seem to be from training data only. Therefore, the relevance of the results is questionable: if, for instance, all the models are overfitting the data, it is not surprising that the performance is similar when training with about 20 % of the data, which of course accelerates training significantly.***

   Response: We have added the information about hyper-parameters.

4. **Comment:** ***The paper's text is difficult to follow due to its disorganized structure and numerous grammatical errors, which disrupt the reader's understanding.***

   Thanks for your comments. We have reorganized the structure of the manuscript and corrected the grammatical error for better understanding.

5. **Comment:** ***A lot of concepts are defined in the introduction and never repeated, e.g. Generative Adversarial Network (GANs), line 48, without reference. Elastic Weight Consolidation (EWC), line 65. Memory Aware Synapses (MAS), line 69. Remanian (do you mean Riemannian?) Walk, line 72. ICARL, line 79. …. What is the relevance of these citations for the proposed method?***

   Response: The proposed incremental learning method is inspired by these referenced methods so that we cite them here.

   Regularization methods involve freezing or normalizing parts of a model when training for successive incremental tasks, preserving knowledge about how to solve different tasks in different parts of the model. Examples include Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017), Memory Aware Synapses (MAS) (Aljundi et al., 2018), and Synaptic Intelligence (SI) (Zenke et al., 2017). These regularization methods focus on penalizing changes to important parameters during incremental task training, and often perform better at alleviating catastrophic forgetting.

Reference

Zenke, F., Poole, B., & Ganguli, S.: Continual learning through synaptic intelligence. International Conference on Machine Learning, 70, 3987-3995, 2017.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., & Grabska-Barwinska, A.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13), 3521-3526, doi:10.1073/pnas.1611835114s, 2017.

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., & Tuytelaars, T.: Memory aware synapses: Learning what (not) to forget. Proceedings of the European Conference on Computer Vision (ECCV), 139-154, doi:10.1007/978-3-030-01219-9_9, 2018.

6. **Comment:** ***Many sentences are not clear, e.g. "The main goal of incremental learning can be described as performing well both in historical tasks.", lines 40-41. In historical tasks and where?***
   Response: Thank you for your good questions. We have rewritten the sentence and it should have been "in historical tasks and incremental tasks".

7. **Comment:** ***"Preliminary conclusion can be drawn from mentioned methods and related literatures is that the similarity/dissimilarity of time series depends on the target of utilizing the similarity, that so far most of the researches propose various measurement methods from time and global or local structural features based on relatively small dataset and that among the methods the most common methods such as Euclidean distance and DTW show high performance with relatively simple idea.", lines 101-105. This phrase is too long, it has some grammatical mistakes, and it is generally difficult to comprehend. "We combine data distribution estimation, temporal similarity, and regularization methods to improve.", lines 142-143. To improve what?***
   Response: Thank you for your insightful question. We have rewritten the sentence into several simpler sentences and corrected the grammatic errors for better understanding.
8. **Comment:** I***"Skewness and Kurtosis are selected as the distribution estimation metrics and standardized Euclidean distance works as the time series similarity metric, the calculation process can be formulated as the following.", lines 182-184.***
   Response: Thank you for your good suggestions. We have rewritten the sentences about the distribution estimation.

9. **Comment:** ***"Then calculating the importance for each parameter in the network is attached to the loss function of the network, as regularization constraint.", lines 198-199. Which is the subject here?***
   Response: Thank you for your insightful comments and constructive suggestions. We have rewritten the sentence for correct expression.

10. **Comment:** "***However, the results show relatively weak self-adaptivity lower the ability of the online learning of the incremental learning method hard to handle the incremental data with rapidly changeable distribution." Lines 323-325***
    Response: Thank you for your good question. We have rewritten the sentence for correct expression.

11. **Comment:** ***Figure 2: for clarity define DNN here (defined on main text at line 2***
Response: Thank you for your valuable comments. We have repainted the figure and given the clear definition of the DNN.

12. **Comment:** ***Figure 3: This figure is quite confusing: how is slicing performed here? Are we measuring the similarity between what? How is new data selected***.
Response: Thank you for your valuable suggestion. We have added the description of the slicing performing process, the object of measurement and the way of data selection.

13. **Comment:** ***Figure 4: FC and R are not defined here and everywhere in the text. Why repeating R if it is a RNN? The picture of the LSTM does not referee the memory cell. Shouldn´t it be C_{t-1} instead of h__{t-1}?***
Response: Thank you for your insightful suggestion. We have checked the picture and formulation of LSTM.

14. **Comment:** ***Figures 5: is this plot relative to the attention-LSTM? And the other DNN models.***
Response: Thank you for your good suggestion. The Figure 5 is about the attention-LSTM, and we have added description for better understanding.

15. **Comment:** ***Line 203: "θθ∗PP,ii is the standard to evaluate the parameter, which represents the difference between the previous and incremental meta sample", what does it mean? What is a meta sample here?***
Response: Thank you for your valuable question. The phrase "incremental meta sample" may be not appropriate here and we have rewritten the sentences to illustrate the parameters.

16. **Comment:** ***Line 208: " When the incremental data come at some time, both baseline and the incremental learning method are performed." Was the baseline not trained once and for all with all the data available?***
Response: Thank you for your good suggestion. Yes.

17. **Comment:** ***Line 210: "…three incremental tasks…". Where are these tasks defined and discussed?***
Response: Thank you for your good suggestion. We have added the description of the incremental tasks.

18. **Comment:** ***Line 270: the NSE is not referenced.***
Response: Thank you for your good suggestion. We've checked all Figures in the manuscript, and added units to the color bars. Besides, we have also adjusted the colormaps for the figures to make them look more scientific and prettier.

19. **Comment:** ***Tables: the tables show only the results for attention-LSTM. Where are the results for attention-GRU and attention-RNN?***
Response: Thank you for your good suggestion. We've checked all figures in the manuscript, and added units to the color bars. Besides, we have also adjusted the colormaps for the figures

to make them look more scientific and prettier. The revised figures, as well as newly added figures, are shown in the following.

20. **Comment:** ***Line 312: do you mean "Good ability on continuous incremental learning?"***
Response: Thank you for your good suggestion. Yes, and we have rewritten the sentences for easier understanding.

21. **Comment:** ***Lines 357-358. The hyper-parameters are not reported, and the results are not robust due to the lack of validation and test splits.***
Response: Thank you for your insightful suggestion. We have added description about the hyper-parameters of the models and the dataset splitting process.

Special thanks for your insightful comments and helpful suggestions on our work. We really appreciate it for it helps us a lot in improving the quality of our manuscript. We have tried our best to make revisions accordingly to improve the manuscript. We hope that the revisions could meet with approval.

Yours,
Sincerely,
Changjiang Xiao