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 1:

1. **Comment:** ***The novelty of the method is not sufficiently described. Currently, it is common practice to train a model on most of the data and then 'finetune', or update the model, using a smaller selection of previously unseen data (see the Neural Hydrology package as an example: https://neuralhydrology.readthedocs.io/en/latest/tutorials/finetuning.html). As this finetuning would currently be the method performed in practise under the situation described here (updating a model during an emergency situation to avoid the time needed for complete retraining), the novelty of this paper appears to be in the selection of the finetuning data. However, the need for the new selection method is not made clear***.

Response: Thank you for your good suggession. We have rephrased the novelty of selection of the finetuning.

2. **Comment:** ***There is a lack of reference to current state-of-the-art rainfall-runoff modelling with RNNs. The rather significant body of work around rainfall-runoff modelling with the types of RNNs used here (LSTMs, GRUs) is not mentioned at all. Nor is the machine learning practice of 'finetuning', which the proposed method is based on, and the use of it in previous rainfall-runoff applications. Citations relating to rainfall-runoff modelling appear to be from the local area, and do not reflect the significant global developments in the field of rainfall-runoff modelling with machine learning.***

   Response: Thank you for your good suggession. This paper focus on the efficiency improvement of the rainfall-runoff modelling with common RNNs. Therefore, it is not the state-of-the-art RNNs but different RNNs in the case of rainfall runoff simulation that were taken into consideration.

3. **Comment:** ***The baseline conditions being compared to are not appropriate. Here, the baseline is an RNN trained with the entire dataset. Whereas, more appropriately, the baseline conditions should be a model that has been trained and then finetuned on a smaller selection of data. This is the current method that would be used in the case of wanting to update a model quickly based on a small amount of newly acquired data, which is the premise of this paper. When comparing the proposed method to a baseline method, the current state-of-the-art (a finetuned model) should be used as baseline. This would then presumably be compared to one finetuned with data selected by the proposed method.***

   Response: Thank you for your comments. Indeed, "baseline" is not appropriate here and "control group" may be better for the description.

4. **Comment:** ***Model setup and training is not performed to currently accepted standards. Hyperparameter tuning, a basic necessity of any machine learning model training procedure, is not performed at all. Instead, values are merely copied from other unrelated studies. In Lines 337 and 358, it is indicated that the study results are demonstrated 'under proper hyperparameter settings' which is apparently not the case. Also, the data appears not to have been split into training, validation and testing sets, to ensure that the reported test metrics are obtained on data that was not used during model training. Data splitting is a staple necessity of machine learning model training to avoid data cross-contamination. If the models are not setup and trained to best practice, why would readers trust the results? There's no indication that the results would hold when readers apply them to rigorously setup and trained models***.

   Response: We have added the details about data splitting into training, validation and testing sets during the model training. The reference is to show the general hyperparameter setting in the similar situation that the models are applied in river rainfall-runoff simulation.

5. **Comment:** ***The stated results are not supported by the reported metrics (that I can tell). The tables of results are visually difficult to comprehend and I am unable to find the stated conclusions within them.***

Response: Thank you for your good suggession. We have revised the stated results form for better understanding and we have changed the tables into figures.

The figure7 presents the results of a comparison between incremental learning and baseline methods on attention-LSTM in terms of root mean square error (RMSE), percentage of error (PE), Nash-Sutcliffe efficiency (NSE), and computation time (TIME) at different stations. The Incremental Tasks refer to simulation tasks of adding equal amounts of new real-time rainfall and runoff data sets divided in chronological order, represented by numbers 1, 2, and 3 respectively. The Method refer to whether using the incremental learning method, represented by baseline and increment respectively. The incremental learning method is seen to have a higher RMSE and lower NSE than the baseline method at all stations, indicating better performance in forecasting. Specifically, the RMSE of the incremental learning method increases by 6.8% to 17.9% compared to the baseline method, with the rise of PE ascend from 1.34% to 2%, while the NSE is reduced vary from 0.02 to 0.06. Moreover, the computation time of the incremental learning method is shorter than that of the baseline method at all stations, with an average reduction to 22.8%. These results suggest that the incremental learning method is effective in improving the efficiency of hydrological forecasting, enabling prompt support for emergency flood prevention and mitigation decisions while maintaining an acceptable range of model training errors. Among the ten stations, the training time of the baseline method and the incremental learning method at the Baihe station is slightly shorter than that of the other stations due to the smaller number of data samples. However, the incremental learning method still shows a consistent pattern of improving training efficiency with the other stations, with an error range of around 1%.

It can be obviously concluded from figure 7 and figure 8 that when the data size is at 20% of the entire dataset, if the model training time increases by more than 4 times, and the difference in error is less than 5%, and the difference in ratio-based metrics is less than 0.08. Consistent results across different attention-RNN models on the same dataset.

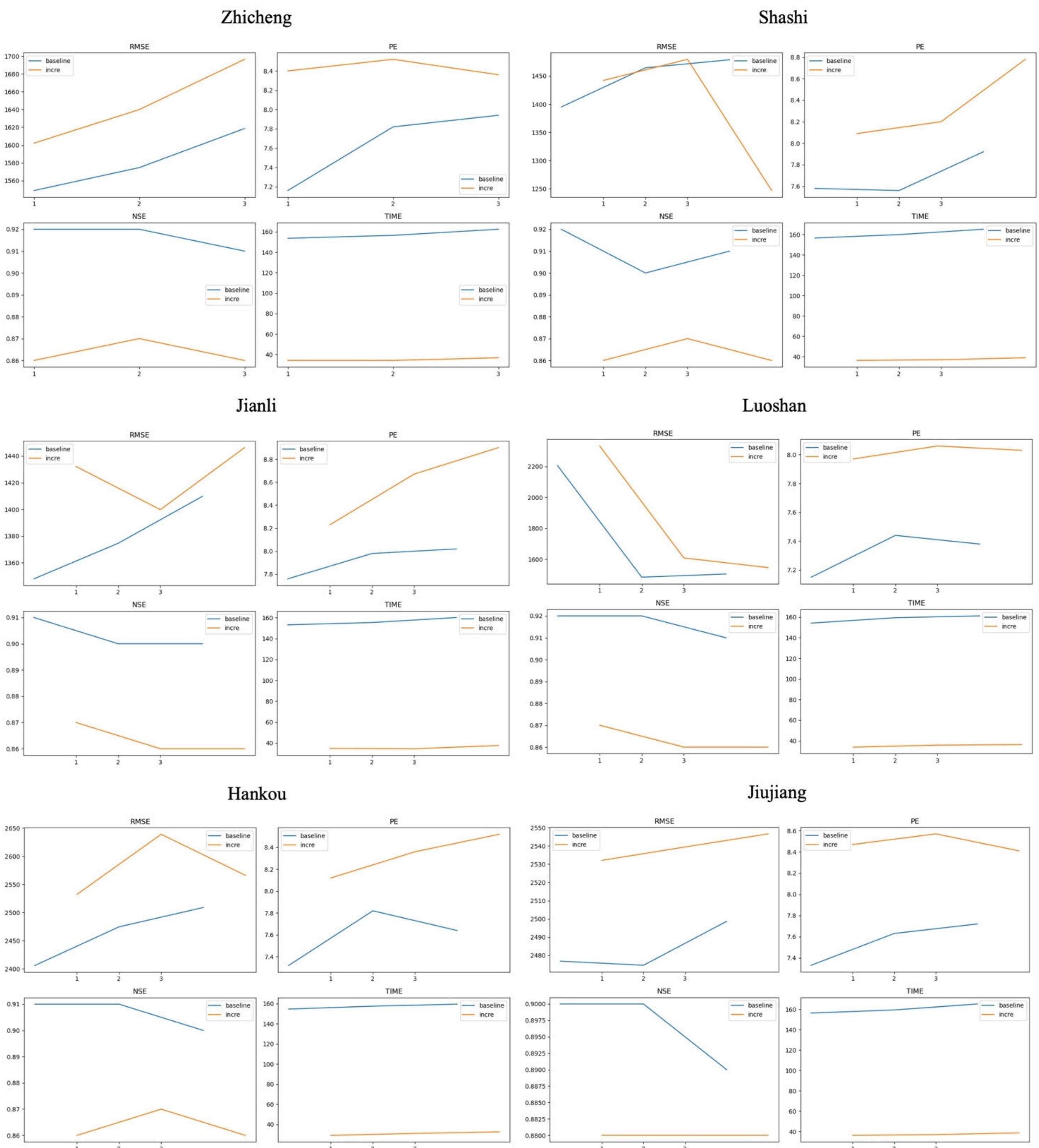

Figure 7: The comparison of performance between the incremental learning method and baseline conducted on attention-LSTM in different stations in Yangtze River basin.

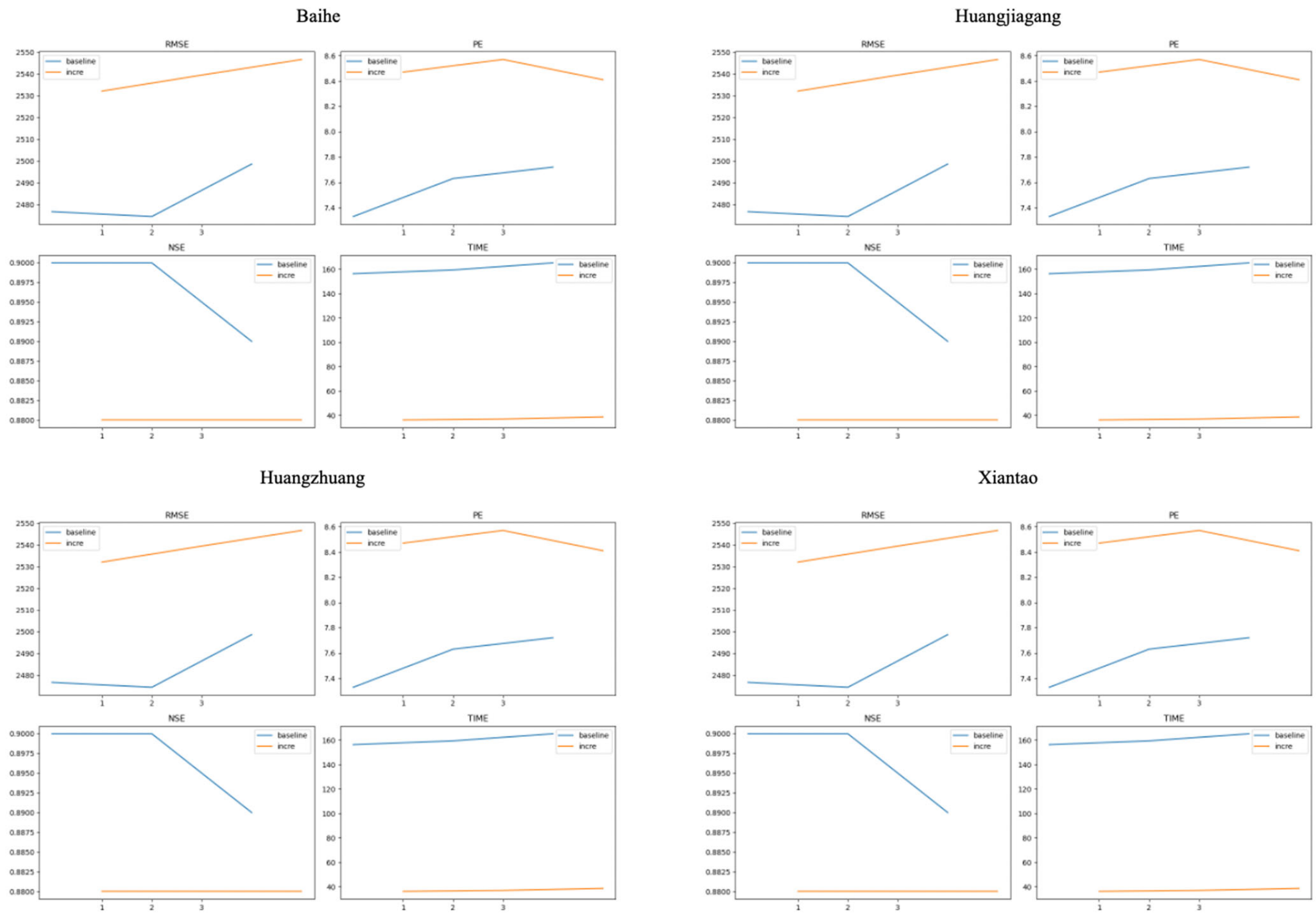

Figure 8: The comparison of performance between the incremental learning method and baseline conducted on attention-LSTMin different stations in Han River basin.

6. **Comment:** ***The overall presentation of the paper is poor. Many sentences are incomprehensible. Confusing terms are used that are not explained and appear to not be related to the proposed method. Much editing is required to ensure sentences are clearly formed and meaningful.***
Response: Thank you for your good suggestion. We have rewritten these sentences to make them easier to understand and explained some ambiguous terms to enhance the readability of the manuscript.

7. **Comment:** ***The basis for this study as mentioned in the first sentence, that 'deep learning always costs plenty of time for training', is too vague. Why would it not be optimal to use a pre-trained model, as is current best practice? The need for the proposed method is not made clear.***
Response: Thank you for your good suggestion. We have clarified the necessity of the proposed method.

8. **Comment:** ***The two-page long single paragraph (obviously far too long) beginning 'Incremental learning…' appears to consist of random sentences from other papers (with appropriate citations given). For example from line 64: 'The reparameterization leads to a factorized rotation of the parameter space and makes the diagonal Fisher information matrix assumption more applicable' - most of these terms have not been used before in this paper and will not be used again. The***

flow does not make sense and much of it seems irrelevant.

Response: Thank you for your good suggestion. We have deleted the redundant expression and explained the principle of the kind of incremental learning method in an appropriate way.

9. **Comment:** ***The description of incremental learning in line 40 as '….learning from…different tasks or domains to solve the future problem with historical experience' sounds like a description of transfer learning. In line 57, 'It is found that neural network required fewer training epochs to reach a target error on a new task after having learned other similar tasks' also describes transfer learning or finetuning, with no reference to either of these well-established machine learning methods. If the proposed method is based on these methods, they should be discussed.***

   Response: Transfer learning and incremental learning have similar connotations, but their focuses are different. This article mainly discusses incremental learning and its application in rainfall and runoff, so there is no need to discuss transfer learning.

10. **Comment:** ***Line 40: 'The main goal of incremental learning can be described as performing well both in historical tasks.***

    Response: Thank you for your good suggestion. We have changed related description.

11. **Comment:** ***Line 46: 'In raw replay methods, a buffer is usually set to store part of the historical data, which avoids frequent data selection when incremental data come while adds memory overhead.***

    Response: Thank you for your good suggestion. We have changed related description.

12. **Comment:** ***Line 66: '…learning is slowed down by weights that are important to the previous task. Specifically, the learning of the important weights that are important to the previous tasks is slowed down.***

    Response: Thank you for your good suggestion. We have changed related description.

13. **Comment:** ***Line 105: 'Owing to the temporal characters of rainfall-runoff data, the similarity measurements for time series can be integrated to partial representative replayed data selection standards of the incremental learning method.' (??)***

    Response: Thank you for your good suggestion. We have changed related description for better understanding.

14. **Comment:** ***Many terms are used in an unclear and unexplained manner, for example: catastrophic forgetting (line 56), important weight (line 69), path integrals (line 73), SI (line 74), ICARL (line 79), etc. These are not explained and the relevance to the paper is not well-defined.***

    Response: Thanks for your comments. We have added the detailed interpretation of the mentioned terms.

15. **Comment:** ***The overall benefit of the method - including historical data in the incremental learning process - is not made clear. Why not just finetune with the new data?***

    Response: Thank you for your good suggestion. Just fine-tuing the new data does not conform to the paradigm of incremental leanring which requires that model take into account the performance on new and old data at the same time.

16. **Comment:** ***A reader could not recreate this experiment given the information here. The method section appears to consist of random sentences pieced together.***
Response: Thank you for your good suggestion. We have reorganized the language to describe the relevant information and process of the experiment.

17. **Comment:** ***Line 154-157: the sentence is long and hard to follow. Should rephrase it.***
Response: Thank you for your good suggestion. We have deleted the redundant expression and explained the principle of the kind of incremental learning method in an appropriate way.

18. **Comment:** ***There is no sufficient explanation of how historical data is combined (ie. Line 165).***
Response: Thank you for your good suggestion. We have given sufficient explanation of the process of historical data.

19. **Comment:** ***Line 164: '…the weights of the calculation result of the difference are assigned and the replay scores are obtained' does not clearly describe how the replay scores are obtained.***
Response: Thank you for your good suggestion. We have added clear description.

20. **Comment:** ***New terms are introduced and not explained: 'depth model parameters', 'moment model', 'incremental meta sample'. These are only used once and never referred to again, increasing confusion.***
Response: Thank you for your good suggestion. We have explained the introduced terms to avoid confusion.

21. **Comment:** ***Line 152: '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.' Why is it a result that the amount of calculation is reduced if using regular network training?***
Response: Thank you for your good suggestion. Here in the manuscript, we justed used the regular network for experiments and observed that the amount of calculation is significantly reduced. In the future, more networks can be tested to see if similar results can be observed.

22. **Comment:** ***Section 2.3, 3, 4: The authors should combine these into an experiment section, and maybe put the discussion of the experiment into a separate section.***
Response: Thank you for your good suggestion. We have adjusted the mentioned chapter structure of the paper for clear and logical elaboration of the experiments of the paper.

23. **Comment:** ***Line 158 refers to this method handling the 'error problem of the network' when this error problem is never described.***
Response: Thank you for your good suggestion. We have added the description of the details of the error problem.

24. **Comment:** ***Figure 2: the image of a feed-forward network used repeatedly here is confusing,***

when the paper is about recurrent neural networks.

Response: Thank you for your good suggestion. We have repictured the Figure 2 for better understanding.

25. **Comment:** ***Figure 4: no mention of what the letters on the diagram refer to (eg. R, FC).***

    Response: Thank you for your good suggestion. We have rewritten the sentences to illustrate the letters. As for R, it's the abbevaration of RNN and the FC means full connection layers.

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

    Three modified RNNs are chosen as the networks. Besides, taken the rainfall and runoff of last station in the upstream as factors, attention mechanism (Vaswani et al., 2017) is applied in RNNs to capture the spatial relationship between the two stations. Inspired by a kind of dual stage attention (Qin et al., 2017), we construct a hybrid network called attention-RNN is shown as Figure 4. attention-RNN consists of four neural network layers, including the encoder attention layer to capture spatial features, the decoder attention layer to capture temporal features, two single-track RNN layers and two full connected layers. RNN units can be regular RNN, LSTM and GRU. FC refers to full connection layer. During the regular training step in the process of the incremental learning method, firstly expand forward calculation according to the time series, and then update the network parameters based on back propagation when three types of RNNs are picked up as the unit to make sure the incremental learning method can perform well in different RNNs. And the network can be specifically named as attention-RNN, attention-LSTM and attention-GRU respectively. Simplified illustration about the three RNNs can be described as following with several formulations.

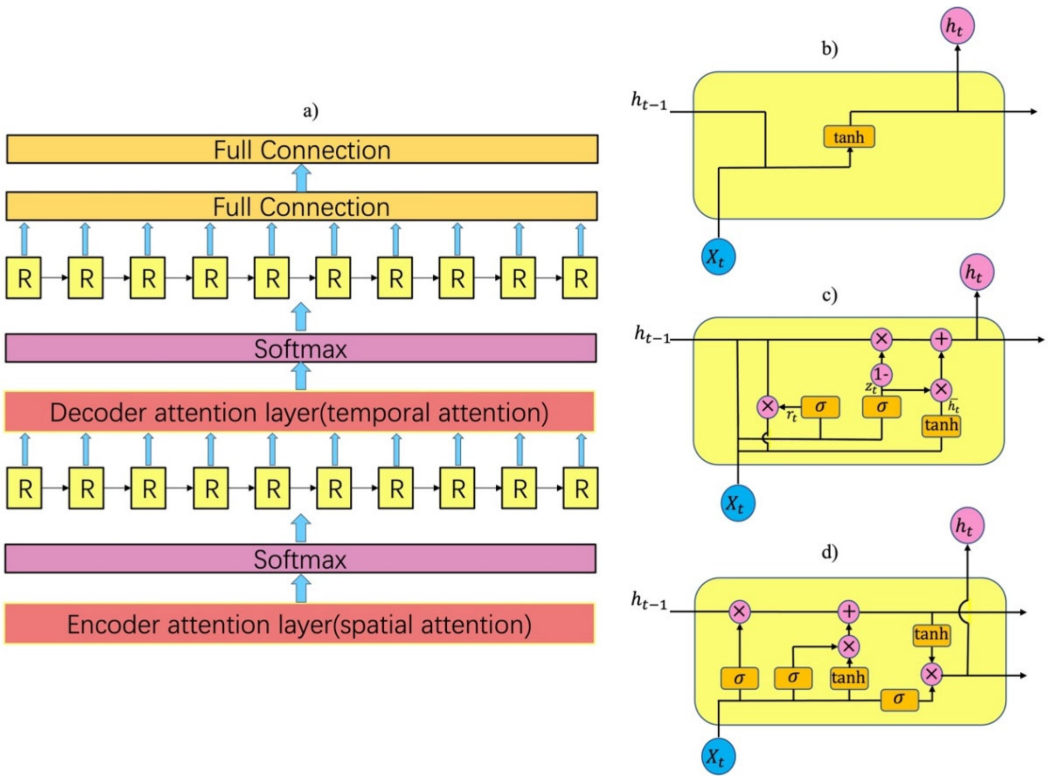

Figure 4: The structure of experiment models and inner structure of different RNNs. a) The structure of attention-RNNs (Qin et al., 2017). b) Inner structure of regular RNN. c) Inner structure of GRU. d) Inner structure of LSTM.

26. **Comment:** ***Time is apparently one of the main metrics of this study and there is no description of how it is calculated (eg. start point, end point, what are the computing conditions, etc.).***

Response: Thank you for your good suggestion. We have added the detailed description of the calculation process.

The related part is illustrated in Section 2.3 Evaluation Metrics.

As for computational time, assuming: total dataset size is N, batch size is B, number of epochs is E, and each epoch requires N/B batches of training, the training time for each batch is T, time optimization algorithm is T_O, the total training time can be described as Eq. (23).

$$Time = \frac{N}{B} * E * T + T_O \tag{23}$$

27. **Comment:** ***Hyperparameter tuning is non-existent. This should be performed with a grid search (or other accepted method) and the choices made should be documented. Borrowing values from other unrelated studies, as described in Lines 282-286 (currently in Results section, should be in Method section), is not adequate. Python packages exist to do this quickly or it is simple to code for yourself. Selection of network size, lookback length and learning rate need to be included.***

Response: Thank you for your good suggestion. the network consists of a total of six RNN layers, with the number of hidden units in both the encoder and decoder set to 128. The RNN step size is set to 7, and the initial learning rate is 0.001. The learning rate is reduced to 0.9 of its previous value every 10000 iterations.

The related part is illustrated in Section 3.1 Experiment Settings.

By referring to the common rules of neural network training and repeated experiments, we choose to set the relative parameters as followings. The weights of distribution estimation metrics and time series metrics are both set to 1. The sample size of each batch or batch size is set to 128 (Gao et al., 2020). The network consists of a total of six RNN layers, with the number of hidden units in both the encoder and decoder set to 128. The RNN step size is set to 7, and the initial learning rate is 0.001. The learning rate is reduced to 0.9 of its previous value every 10000 iterations.The mean square error (MSE) is selected as the original loss function for the gradient descent method to the baseline. And regularization constraint is added to MSE decrease in the incremental learning method. The optimization algorithm is Adam. The regularization parameter λ is set to 1 (Aljundi et al., 2018).

Reference

Gao, S., Huang, Y., Zhang, S., Han, J., Wang, G., Zhang, M., & Lin, Q.: Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation. Journal of Hydrology, 589, 125188, doi:10.1016/j.jhydrol.2020.125188, 2020.

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.

28. **Comment:** ***Line 158 refers to this method handling the 'error problem of the network' when this error problem is never described.***

Response: Thank you for your good question. Yes, lower RMSE and higher NSE indicate better performance. And we have added the relation between the numerical value and performance.

29. **Comment:** ***Line 294: 'the RMSE of the incremental learning method increases by 6.8% to 17.9% compared to the baseline method' followed by Line 297: 'These results suggest that the incremental learning method is effective in improving the efficiency of hydrological forecasting… while maintaining an acceptable range of model training errors.' How is it determined that an increase of 17.9% is within an acceptable range? In Line 344 this is referred to as a 'smaller error compared to the baseline model'. This needs more explanation.***

    Response: Thank you for your good suggestion. We have added the description about the calculation and meaning of the numerical value.

30. **Comment:** ***Line 302: '….with an error range of around 1%.' There is no description of how this 1% is calculated or what it means.***

    Response: Thank you for your good suggestion. We have added the description about the calculation and meaning of the numerical value.

31. **Comment:** ***Line 310: 'It can be obviously concluded from table 1 and table 2 that when the data size is at 20% of the entire dataset, if the model training time increases by more than 4 times, and the difference in error is less than 5%, and the difference in ratio-based metrics is less than 0.08.' This sentence is unclear. I do not know what the 20% refers to, and it is very difficult to find the 4 times increase and difference in error (what error?) of 5% on the tables. What is ratio-based metrics – these have not been described and it is unclear how the value of 0.08 has been determined?***

    Response: Thank you for your good suggestion. We have added the description about the calculation and meaning of the numerical value.

32. ***Line 323: 'Specifically, the run-time difference reaches over 4 times, the PE increase less than 3%, the NSE decease less than 0.05.' Again, support for this is claim cannot be found in the results.***

    Response: Thank you for your good suggestion. We have rewritten the sentences to describe the results.

33. **Comment:** ***Line 344: 'However, it is notable that the baseline model and the incremental learning method had a higher error in the Han River basin than in the Yangtze River basin, likely due to the similar climatic conditions and rainfall patterns between the two regions.' If conditions are similar, why are errors expected to be different?***

    Response: Thank you for your good suggestion. The different errors derive from the different temporal characteristics of the rainfall-runoff data between Yangtze River Basin and Han River Basin. Since the absolute value of the runoff of the Han River is small, it is more obviously affected by rainfall. There is a large gap between the wet season and the dry season, and it is more volatile in terms of time series characteristics. After selecting representative samples, the effect on incremental tasks is poorer, while the absolute value of the Yangtze River runoff is large and the time series characteristics are more stable, the effect of the incremental learning method will be slightly better. This article has discussed this issue in the section on the applicability of incremental learning methods to watersheds. You can view the relevant statements in Section 3.4. For better understanding, we have revised related discussion.

34. **Comment:** ***Line 302: '….with an error range of around 1%.' There is no description of how this 1% is calculated or what it means.***
Response: Thank you for your good suggestion. We have rewritten the sentences to illustrate the meaning of the metrics.

35. **Comment:** ***Figure 5: what are the units on the y-axis of the Time plot? These results appear suspiciously close together, in the range [4.1-4.5], for all of the stations and all of the models. How is this explained?***
Response: Thank you for your good question. We have revised the pictures. The results are not totally the same. The data of adjacent stations are relatively close, and the fitting effects of the three models constructed in this article are relatively close under the same hyperparameters. However, what the paper concentrate on is that the proposed method's function to the different models on different stations. The results are intended to illustrate the suitability of the proposed method.

36. **Comment:** ***Figures 6 and 7 should be combined into one figure.***
Response: Thank you for your good suggestion. In order to better discuss the characteristics of the methods in different basins, we insist on putting them on two figures.

37. **Comment:** ***Many phrases are incomprehensible, for example Line 329: '…which imply that when the incremental data are taken as continuously input, the incremental learning method gives the deep learning models the ability to continuous incremental learning.' and Line 360: 'Besides, the similar increase intensity of evaluation metrics differences shows that….'***
Response: Thank you for your good suggestion. We have rewritten these phrases for butter comprehension.

38. **Comment:** ***Again, terms are used that are not described and are not used again, eg.: 'distribution rules', 'weak self-adaptivity', etc.***
Response: Thank you for your good suggestion. We have described the terms for better understanding.

39. **Comment:** ***The three listed conclusions are unclear and unsupported. In the second point, the claim that the proposed method '…guarantee percentage error increase and NSE decrease less than 5%' has not been clearly demonstrated.***
Response: Thank you for your good suggestion. We have reclarified the related description to make the conclusions more clear and more supported.

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