Author Response to Review of

How can regime characteristics of catchments help in training of local and regional LSTM-based runoff models?

Reyhaneh Hashemi, Pierre Brigode, Pierre-André Garambois, Pierre Javelle

HESS, doi:10.5194/hess-2021-511

RC: Reviewer Comment AR: Author Response

Reviewer: Anonymous Referee #2

AR: We would like to thank the Anonymous Referee #2 for their review and constructive thoughts, which we believe will help significantly improve our paper. The reviewer is critical of the design of our regional experiments and identified technical issues warranting a revision of these experiments. We acknowledge certain limitations and caveats regarding this aspect and agree with many of the points the reviewer makes. We plan to make the necessary revision(s), given further below, to address these issues. It seems to us that few points regarding details of our experiments are misunderstood. These points are clarified where they are elaborated on. We would like to invite the reviewer to find our point by point response to their individual comments in the following sections.

1. SUMMARY OF REVIEW

RC: This paper addresses two research questions related to the use of LSTMs for rainfall-runoff modeling: (1) Does appropriate sequence length depend on hydrological regime, and (2) should LSTM training be done on hydrologically similar basins?
To state my opinion up front, I have run similar experiments (unpublished) and found results that are qualitatively different than what are reported here. There are several technical issues in this paper (overall, the methodology is not appropriate for testing the stated hypotheses), and it might be worth addressing those before we look carefully at the results.

AR: We thank the reviewer for their interest in performing similar experiments. We would be happy to engage in an ongoing dialogue with the reviewer about the details of their experiments since without further information, in particular, on the hydro-geo-climatic context of their data, it would be hard to provide a definite explanation. Indeed, such discrepancies could be investigated at different levels. At the highest level, we would conjecture that the reason lies in definition of the homogeneity component. We believe that the hydrological similarity rule — i.e., the regime classification — is a crucial question. Given the term "similar experiments", we could think of the following two cases, which both could have "very" different learned latent spaces compared with ours.

Case 1 The exact same classification is applied to a sample in a non French context. In this case, we would be afraid that the exact same rule would not be immediately applicable to other climatic contexts.

The following elaboration on our classification approach aims to underline how it is intensely context dependent — in terms of number of classes, criteria, and thresholds.

As a property of the French context, we knew in advance that there existed five main regime patterns, which we named Uniform, Mediterranean, Oceanic, Pluvial, and Pluvial-Nival. Therefore, any catchment in our sample could be classified in one of the five categories. Using the fact sheets available at `https://webgr.inrae.fr/activites/base-de-donnees/`, we tried to identify hydro-geo-climatic signals that could reflect different features of all five patterns. The decision feature(s) — based on which we could distinguish one pattern from the others — was (were) not the same for all regimes. For instance, we were observing that the minimum temperature attribute alone was able to detect the Nival pattern. While, in catchments with known water ground effects, it was certainly a criterion on discharge that was doing this. In the same spirit of a decision tree algorithm, but at a human level, we concluded that the mean annual discharge, total precipitation, and temperature signals would be the most useful signals to exploit to identify the distinctive attribute(s) in each class. After a number of trial and errors on different properties of these signals, such as the number, magnitude, and time of occurrence of their global/local peaks, we identified our classification attributes — IQ, IP, and $T_{\min}$ — and their thresholds. Therefore, inferring the similarity rule for any other climatic context warrants a redefinition of different elements in this analysis.

Case 2 Another classification (concluded at an AI or a human level) is applied to data belonging to a non French context. In this case, getting different results would not be surprising — since a different rule involves different decision attributes and criteria. The question in this case would be rather the extent to which we could compare the results obtained from two different classifications.

To conclude on this point, we believe that such cross study comparisons warrant caution, since the imposed similarity rule will not be identical between the studies. That is why we would like to emphasis and acknowledge that research questions established on a subjective component, such as regime classification in our paper, will always make the corresponding conclusions subject to that component. The suggestion we made in Section 6 of the paper aimed to address this point:

> The current conclusions are drawn in the French climatic context. Validating these conclusions in a different climatic context (e.g. using the available CAMELS data for the US, Great Britain, or Brazil) would help to suggest more widely applicable training approaches.

RC: My overall recommendation is to revise the experiment as suggested in one of the comments below. The experimental design that is appropriate to test the (two) hypotheses outlined here is very simple (but somewhat computationally expensive). If the authors were to find similar results using a more appropriate experiment, this would be an interesting study.

AR: We provide our response to this comment later in Subsection 2.3 where the reviewer details their suggested design of experiments.

2. COMMENTS

2.1. Hyperparameter tuning

RC: Hyperparameter tuning was done on LSTMs trained on individual basins. LSTMs trained on individual basins behave fundamentally differently than LSTMs trained on multiple basins, which means that lessons learned from hypertuning on individual basins do not translate to multiple-basin models.

AR: The reviewer identifies two unfavorable points:

Point 1 "LSTMs trained on individual basins behave fundamentally differently than LSTMs trained on multiple basins".
We would assume that the reviewer's term "LSTM's behavior" could be translated to "LSTM's performance measured by the median KGE". Taking into account each individual panel of Figure 10 [1] and comparing the median KGE of its local LSTM (the leftmost pair of box plots) and the regional LSTMs (the box plots in the center), we do not recognize any fundamental difference in their performances. Except for the Mediterranean panel, the local LSTM and at least one regional LSTM have a very similar median KGE.

Point 2 "[...] which means that lessons learned from hypertuning on individual basins do not translate to multiple-basin models".
As the reviewer confirms in their next comment, the only benefits we got from the hyperparameter tuning experiments were deciding to use what model structure with what batch size. Furthermore, we believe that we do have no evidence to confirm or reject the utility of these two tunings for regional training — we simply did not test any other variations of them (batch size and model structure) for regional LSTMs.

RC: Additionally, 15 catchments is not enough for robust hypertuning – we would need to perform hyperparameter tuning on the full (evaluation) dataset (although see a later comment – the experimental design needs to be changed fundamentally). Also, notice that the only portions of the "hypertuning" that were actually used for the other experiments in this paper were (1) discarding the S2 model architecture, and (2) batch size.

AR: (For the sake of convention, please let us use the terms "optimal" and "tuned" interchangeably hereafter.) We completely agree with the reviewer's point that a robust hyperparameter tuning warrants a sample larger than 15 when the population size is over 700. Our reading of this is that in order to "maximize" the benefits of hyperparameter tuning over the entire population, more than 15 catchments should be taken. This means that: to get close to the optimal point of as many of catchments as possible we need to perform tuning on more than 15 catchments. This is a sound argument. We therefore would like to clarify that our goal was not to be tuned with respect to every single point of the population. Our goal was rather to be in a broadly useful zone in the space of hyperparameters. We are aware that such zone would be close to the optimal point for some instances, far away from the tuned point for some others, and possibly at the exact same position of the optimal point for some (very rare) cases.

Above all, please note that the way that hyperparameter tuning is performed in this paper does not allow — on its own — to be carried out for all catchments. The reason is that all the results presented for our hyperparameter tuning, and therefore all the conclusions drawn form them, correspond to the evaluation period (last interval). This indicates that their evaluation data are already used in hyperparameter tuning and no unseen data are left for these catchments. That is why they need to be immediately excluded from the sample. If we did hyperparameter tuning on all catchments, we should have excluded them all — already at the hyperparameter tuning stage — and we would have had no more study catchments for our main experiments (EXP1 and EXP2). The reviewer might ask why we adopted such tuning approach. The answer is since in the main experiments models are being evaluated on the evaluation period and we believe that in hyperparameter tuning we need to do an identical evaluation.

[1] in the Preprint version

We would nevertheless like to state that we have no objection to drop the section (and any other contents) related to hyperparameter tuning, if the reviewer finds it to be superfluous.

2.2. Number of hidden units in the LSTM layer

RC: There is strong relationship between the dimension of the cell state and the sequence length, and also between the cell state dimension and the ability of the model to generalize (Kratzert et al. (2019) shows how the model uses the cell state to map catchment similarity). This parameter was not included in the hyperparameter tuning, and it was also not considered in the experimental design. 64 cell states is smaller than used by most of the previously published work. The hypotheses that are tested here are about the ability of the model to generalize and about memory timescales, both of which are directly controlled by the cell state (more cell states means more ability to have different memory timescales for different hydrological regimes).

AR: We thank the reviewer for this relevant and constructive comment regarding the number of hidden units in the LSTM layer. We also acknowledge the findings of Kratzert et al. (2019) on the relation between this parameter and model generalization ability. We believe that improving the generalization aspect, i.e. the ability of the model perform well on "unseen" data, would be of central importance/interest when applying the regional model to ungauged catchments — catchments not used in model training. Our paper did not intend to investigate any such applications. Besides, in Lees et al. (2021), they use a cell state size of 64 and the Entity-Aware-LSTM model — the model used and developed by Kratzert et al. (2019). We quote from the Preprint version of their paper (available at `https://doi.org/10.5194/hess-2021-127`):

> "We chose the hyper-parameters (dropout rate, hidden size - hs) based on the choices in previous studies (Kratzert et al., 2019)."

and in the final version (available at `https://doi.org/10.5194/hess-25-5517-2021`) they update/add that:

> "We chose the hyper-parameters (dropout rate, hidden size – hs) based on analysis of the NSE performances, finding that the improvement of further model complexity (increased hidden size) was negligible after a hidden size of 64. The hidden size was also consistent with the choices made in previous studies (Kratzert et al., 2019)."

It would be therefore arguable that the most effective hidden unit number would be always much higher than 64 since Lees et al. (2021) reported that this rule did not hold true in their work.

Please note that the preprint version of Lees et al. (2021) was published after Kratzert et al. (2019) and available at the time of preparation of our work. We based explicitly our choice of cell state size on this specific study, which was (at the time) the most recent study.

2.3. Design of experiments

RC: It would be interesting (and useful) to know whether there is value in clustering catchments prior to training models, and if so whether we could find correlations between different hyperparameters (e.g., sequence length, cell state dimension) and hydrological regime (the former is a more interesting question than the latter, in my opinion). The way to test this is simple – you separately (and fully) hypertune each model. For example, if you want to test the clustering strategy described in lines 120-125, you would hypertune models separately for each catchment group (considering all of the important LSTM hyperparameters), and as a benchmark you would hypertune a model for all of the

catchments combined. Then the results would be directly comparable. After that, you could look at whether there was any relationship between hydrological regime and the "optimal" (hypertuning is never actually optimal) sequence length for that cluster. If you really wanted to train single-basin models (which I suggest you should not do), then these need to be separately (and fully) hypertuned for each basin.

AR: We appreciate the detailed description of the suggested design of experiments (DOE) and we acknowledge that it is thorough. We recognize that the reviewer identifies the following limitations for our DOE with respect to our research questions:

Limitation 1 In the first research question, the proposed hypothesis is that the effective size of lookback and regime of catchments are correlated. In order for the hypothesis to be valid, all lookbacks should have been tested for all regimes as well as the entire sample (B_1). This has been done only in local training and not for regional LSTMs.

Limitation 2 In the second research question, the hypothesis is that training less but hydrologically homogeneous catchments is more effective than training more catchments that are hydrologically heterogeneous. In our regional experiments, we have used lookbacks that were concluded at the local scale. Therefore, we might have not taken the most effective lookback for regional models when answering to the second research question.

Limitation 3 In the reviewer's opinion, the number of hidden units should have been varied along with the lookback size.

We plan to revise our DOE as follows:

Revision 1 For all regional — either per regime or per sample — models, we will test all the lookbacks tested for local models. This revision will deal with both Limitation 1 and Limitation 2.

In making this revision, we will consider 64 hidden units — this is necessary since in all local LSTMs 64 hidden units is used.

Revision 2 (Our answer to Limitation 3 has been provided above in Subsection 2.2. This revision is therefore provisional — it might be considered or not.) We will repeat Revision 1 using 256 hidden units. This size was concluded as most effective in Kratzert et al. (2019).

2.4. Interest of local models

RC: I wonder why we are training local models. There is no situation where we would ever use a model trained on a single catchment for any real-world purpose. Additionally, the behavior of the LSTM is fundamentally and qualitatively different when trained on one catchment vs. many, which means that we cannot learn anything general or useful from locally trained models. If there was a specific hypothesis that we wanted to test that required training local models, then this might make sense, but I do not believe this is the case here – we could ask the question about appropriate sequence length on hydrologically grouped models, and asking the question this way would give us a more useful answer. Just a note: Kratzert et al. (in all papers after their 2018 paper) trained single-basin models only to make the point that this is not an appropriate thing to do.

AR: We understand and share the reviewer's interest in using universal regional models. We also acknowledge the work of Kratzert et al. in developing regional models and their potential in transferring knowledge from a number of known situations (gauged catchments) to unknown cases (ungauged catchments). There are however a number of reasons for which we disagree with the opinion that, in general, there is no interest in training local models, and, in particular, it has been done out of the purpose of the paper. These reasons from both general and particular perspectives are discussed in the following paragraphs.

General - 1 Compared to local models trained on individual catchments, regional models are expensive to develop and maintain — the slightest change in their setup necessitates re-training a huge model, even if we do not start from scratch. We agree that such computational expenses would not be a worry for Deep Learning researchers having access to the latest generation of GPUs. This would not be however the case for operational bodies from which real world applications arise. In particular, the interest of many of these operational entities (in France, but presumably elsewhere) is limited to a single (or very few) water course(s).

General - 2 The issue of computational cost of regional training is tied to another question: availability of national databases, which is often not the case — at least to public. In France, the SAFRAN data base used in this study has been made available only to very few research institutes. That is why we believe that our data set constitutes one of the novelty aspects of the paper noting that most of previous studies have been conducted using a single publicly available data set.

We therefore find it idealistic to think that none of these difficulties — i.e. availability of GPU resources and national data sets — will never emerge in practical applications.

General - 3 There are real world cases of catchment for which it would not be appropriate to use universal models. Dam influenced catchments are one of these cases. These catchments do not have natural responses and their non natural response is "unique" to them. We showed in our paper that in, even highly, influenced catchments, LSTM was able to learn the non natural response rule in these catchments. Such rules do not however constitute "transferable" knowledge since for each catchment is a function of the rule of dam, which itself depends on the dam purpose (flood control, irrigation, hydroelectric, ...), and the natural response rule of the catchment.

Paper - 1 In the first research question a general hypothesis is made — there is a link between the LSTM's lookback size and hydrological regime of catchments and it has not been specified that it would hold true solely for regional LSTMs. We believe that this hypothesis needed to be investigated at both scales in order to delimit its boundaries of validity.

Paper - 2 Local trainings in this paper are not carried out in the same way as any of the previous studies. Therefore, it would be arguable to immediately rule them out based on previous studies — having a basically different training approach — and without further investigation. Indeed, we used 10 years of data for validation, 10 other years for evaluation, and the size of train data in our sample varied between 10 and 40 years. While, for instance, in Kratzert et al. (2018) the local LSTM is trained using only 15 years of data and no data is used for validation. Or in Lees et al. (2021), although a validation period is taken into account, it is limited to 5 years and the length of train data does not go beyond 11 years while in our sample we have catchments with train data as long 40 years. (This point has been addressed in the paper, please see Section 5.2 of the preprint version). Also, train duration in our work was not imposed as a predefined fixed number of epochs. Furthermore, the GR4J model benefits from a particular feature — dealing with water gain/loss — making a clear distinction between the "baseline performances" in our work compared to previous studies. Taken all together, as we indicated in the conclusions of the paper (Line 507), we do not agree that regional LSTMs outperform, in any cases, local LSTMs

> and all conceptual models. In particular, our results show that when train data are abundant, validation data are also present and sufficiently long, and train duration is not strictly imposed, local LSTMs could catch up with regional LSTM's level.

We find it thus arguable to state that regional LSTMs should be taken as the default choice in all cases of data and application and for all purposes.

3. MINOR COMMENTS

3.1. S2 architecture

RC: The S2 architecture (stacked LSTMs) is interesting, but not related to either of the hypotheses of the study. What was the motivation for testing this and how does it relate to the questions that were motivated in the introduction? I'm not saying to remove it, just give us some reasoning or motivation. Also, when the "complexity" of this model is discussed, you might give us the number of free parameters so that we can get a sense of what the differences are.

AR: Thank you for this interesting inquiry. The original intent behind studying the S2 architecture was to act on classical instructions that are given for training any Deep Learning models: prevent underfitting. We wanted to see if vertically stacking LSTMs could bring immediately a better performance thanks to a better hierarchical feature learning — in the same spirit of operating successive convolutions in a Convolutional Neural Network if we could "metaphorically" think that what LSTM looks for in time is comparable to what a CNN detects in space. But, we did not observe any instant improvement. There were thus two speculations: either, we still underfit a lot, or, the stacked setup would not basically help. Finding a definite answer to the question of "still underfitting" required an unmanageable amount of work — stacking more LSTMs and increasing hidden units until we overfit and repeating all local experiments at every step — and after all such effort, there was still the risk that "the stacked setup would not basically help". We therefore chose to assume that it was the second hypothesis that held true — stacking LSTMs vertically would not bring significant performance improvement. Thus, we ruled out the S2 structure.

Regarding the term "complexity" in the context of Neural Networks (NNs), the number of trainable parameters might not be a fully representative measure of model complexity in NNs since they are usually over parametrized. Indeed, one could find two different definitions of model complexity in the literature: 1) model expressive capacity (Bengio and Delalleau, 2011; Poggio et al., 2017) or 2) model usable capacity (Novak et al., 2018; Hanin and Rolnick, 2019). According to these definitions, different contributing factors are identified. For instance, choice of activation function, optimization algorithm as well as model size, which could be measured by the number of learnable parameters, the width, or the depth of the network. What we meant by the term "complex" in the paper was in terms of the depth of the network (its hierarchical representation).

We therefore completely agree with the reviewer that it should be specified what measure we are referring to when using the term "more complex" — in particular, when the S2 architecture involves less learnable parameters (13473) than the S1 (18497). We will update this sentence in the new version.

3.2. Purpose of validation set (Line 192)

RC: *The validation set is intended to be used for finding the best weights and biases during training and control overfitting.* I think this is just a typo.

AR: From this comment, we would infer that our hyperparameter tuning methodology is misunderstood — there exists no typo in the marked sentence. As we explained above, in our response to one of the earlier comments of the reviewer, we based our choices of batch size and model architecture on the results obtained using evaluation data and not validation data. Please refer to Subsection 2.1 of the present document.

RC: Validation data is used to help find the best *hyperparameters* and control overfitting (it is explicitly *not* used to help tune weights and biases, except through early stopping).

AR: There is no typo/error in this phrase, neither. There is a missing explanation about how we chose model's best parameters. We did this in the same way as the `restore_best_weights` works in Keras. During training, for each epoch, we stored model parameters at the end of the epoch along with its validation loss. In the end of training, we identified the epoch corresponding to the best validation loss (smallest MSE). We then loaded the model with the parameters corresponding to this epoch and used this specific update of the model to make predictions on evaluation period (box plots with solid edges and darker colors in the paper's figures), as well as train + validation period (box plots with dashed edges and brighter colors in the paper's figures). Validation data are thus explicitly (or implicitly, depending on your definition in the current context) used to conclude on the best model parameters.

We will include this missing explanation in the revised version and we thank you for noticing this.

3.3. Catchments with very short train data (Line 201)

RC: *What remains constitutes the train period (P1) the length of which varies between 1 year to 40 years in the FR sample.* It is a little concerning to have different sized training data records per catchment, especially if some catchments only have 1 year of training data. This is *especially* problematic if we are looking at differences between what data is required to train in different types of catchments.

AR: Please note that "all" catchments with less than 10 years of train data were excluded from our analyses at a very early stage. Please refer to Section 4.1 (Line 337) of the paper.

Please also note that none of the conclusions reported in our paper is based on the results obtained from such instances.

3.4. Training duration (Line 180)

RC: In line 180 is reads like you are doing sequence-to-one prediction, however in line 259 you say that you are using a patience of 50 epochs with a maximum of 500 epochs. Typically you only need this many epochs if you are doing sequence-to-sequence training. Regardless, the number of epochs used by previous studies was in the range of 20-50. Have you found that more epochs help (we looked at this carefully in previous studies), or is there something else about your model that is different from previously published work?

AR: Thank you for this interesting inquiry. One reason that we opted for the early stopping algorithm was that it would not impose to all catchments/simulations a wall condition — the same predefined non traversable training duration. Instead, it allows the model to continue to learn as long as its performance (on the validation set) is improving.

The so large numbers that we considered for these two parameters were meant to provide a free boundary condition. This would give the model the freedom to learn as long as it needed (unknown to us) without being stopped too early — due to either a too little patience or a too small maximum number of epochs.

This feature was advantageous to us since our data set was novel and had not been used in any of the previous studies notably that it contained catchments with very long sequences (up to 40 years).

The training duration — the epoch at which training has stopped — is available for all of our simulations and will be provided upon request.

3.5. Static attributes (Line 291)

RC: This is a pretty small list of catchment attributes. Given that catchment attributes are available globally (e.g., HydroAtlas), and this will directly influence the generalizability of a model, why did we use such a limited set of attributes here?

AR: Thank you for this relevant question. We agree with the reviewer that augmenting catchment specific data helps make local to regional mapping more effective, which in return reduces the generalization error. We believe however that not considering plenty of static attributes would be concerning if the model is being applied to ungauged catchments, which is not the case in our paper. We therefore took the classical static descriptors often used in previous regionalization studies in the French context (Oudin et al., 2008).

3.6. Naming strategy

RC: In general, naming experiments with non-descriptive names like R1, R2. P1, etc. makes the paper more difficult to read than is necessary. This means that the reader must always refer back to the text in order to understand each figure. This can be solved simply by naming each of the models/experiments/datasets with descriptive names.

AR: Thank you for this suggestion. Name revision will be considered to address this issue in all such instances.

REFERENCES

Y. Bengio and O. Delalleau. On the expressive power of deep architectures. In *International conference on algorithmic learning theory*, pages 18–36. Springer, 2011.

B. Hanin and D. Rolnick. Complexity of linear regions in deep networks. In *International Conference on Machine Learning*, pages 2596–2604. PMLR, 2019.

F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger. Rainfall–runoff modelling using long short-term memory (lstm) networks. *Hydrology and Earth System Sciences*, 22(11):6005–6022, 2018. 10.5194/hess-2021-51110.5194/hess-22-6005-2018. URL https://hess.copernicus.org/articles/22/6005/2018/.

F. Kratzert, D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter, and G. Nearing. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, 23(12):5089–5110, 2019. 10.5194/hess-2021-51110.5194/hess-23-5089-2019. URL https://hess.copernicus.org/articles/23/5089/2019/.

T. Lees, M. Buechel, B. Anderson, L. Slater, S. Reece, G. Coxon, and S. J. Dadson. Benchmarking data-driven rainfall–runoff models in great britain: a comparison of long short-term memory (lstm)-based models with four lumped conceptual models. *Hydrology and Earth System Sciences*, 25(10):5517–5534, 2021. 10.5194/hess-2021-51110.5194/hess-25-5517-2021. URL https://hess.copernicus.org/articles/25/5517/2021/.

R. Novak, Y. Bahri, D. A Abolafia, J. Pennington, and J. Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018.

L. Oudin, V. Andréassian, C. Perrin, C. Michel, and N. Le Moine. Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 french catchments. *Water Resources Research*, 44(3), 2008.

T. Poggio, K. Kawaguchi, Q.i Liao, B. Miranda, L. Rosasco, X. Boix, J. Hidary, and H. Mhaskar. Theory of deep learning iii: explaining the non-overfitting puzzle. *arXiv preprint arXiv:1801.00173*, 2017.