Reviewer #1

This paper deals with a highly current subject matter. It presents a methodological framework for streamflow prediction in ungauged basins using a leave one out cross-validation approach (LOOCV) for three hydrological models and an LSTM network. The authors compare the performance of these four models for 148 basins in Northeast North America. The evaluation of the models at such a scale and working with these number of watersheds is quite impressive. The paper is well written, illustrated and organized. However, some points need to be taken into consideration to improve the manuscript, namely:

We would like to thank Reviewer #1 for their comments and suggestions. Here are the point-by-point responses on how we plan on modifying the manuscript in the next version.

(i) A more comprehensive literature review should be provided; that is, the authors should acknowledge the works of several other authors who have dealt with streamflow prediction using LSTMs, not just focusing on the works of a specific research group. Please homogenize the diversity of the literature review and cover the works of others that have provided significant achievements in the field of hydrological modelling using Deep Learning (DL) models.

This is a good point. We will explore other studies and applications and contextualize the work better in this regard. We focused initially on North American studies that applied LSTM models for the regional aspect, but indeed it would be better to also put this into context of other studies from other groups and in other regions as well. We recommend adding the following works and describing their findings in the revised version of the manuscript:

Ayzel G, Kurochkina L, Abramov D, Zhuravlev S. Development of a Regional Gridded Runoff Dataset Using Long Short-Term Memory (LSTM) Networks. Hydrology. 2021; 8(1):6. https://doi.org/10.3390/hydrology8010006

Ayzel, G., Kurochkina, L., Kazakov, E., & Zhuravlev, S. (2020). Streamflow prediction in ungauged basins: benchmarking the efficiency of deep learning. In E3S Web of Conferences (Vol. 163, p. 01001). EDP Sciences.

Choi, J., Lee, J., & Kim, S. (2022). Utilization of the Long Short-Term Memory network for predicting streamflow in ungauged basins in Korea. Ecological Engineering, 182, 106699.

Li, X., Khandelwal, A., Jia, X., Cutler, K., Ghosh, R., Renganathan, A., Xu, S., Tayal, K., Nieber, J., Duffy, C. and Steinbach, M., 2022. Regionalization in a global hydrologic deep learning model: from physical descriptors to random vectors. Water Resources Research, 58(8), p.e2021WR031794.

Nogueira Filho, F. J. M., Souza Filho, F. D. A., Porto, V. C., Vieira Rocha, R., Sousa Estácio, Á. B., & Martins, E. S. P. R. (2022). Deep Learning for Streamflow Regionalization for Ungauged Basins: Application of Long-Short-Term-Memory Cells in Semiarid Regions. Water, 14(9), 1318.

(ii) In the proposed LOOCV for LSTM modeling, the LSTM model was trained using a large dataset (N-1 basins), while keeping one basin as a pseudo-ungauged basin for validation. This approach departs from the basic philosophy of training DL models. Indeed, to avoid introducing a bias during training of DL models such as LSTM, overfitting should be avoided by considering a considerable proportion of the whole dataset as a testing dataset. What was the rationale behind this methodological approach?

Actually, this is incorrect, but we acknowledge where the confusion comes from. We forgot to clearly state the percentage of catchments used in training and validation in the methods and only refer to the values in line 408 of the original manuscript, where we explain that it is 80% of (N-1) basins used in training and 20% used in validation, with the pseudo-ungauged basin being the only one used as the "testing" basin. This is to ensure the LSTM model does not over-train/overfit, and the "regionalization" skill is evaluated on the completely independent testing basin. This will be clarified in the methods to ensure readers get the whole picture before the end of the results section.

(iii) The authors propose an LSTM modeling approach for ungauged basins that will, without a doubt, spur the interest of the readers. However, the literature has provided several good performances of LSTM models for similar regions in Northeastern North America. Perhaps the authors could provide some insights for future work in dry regions where the presence of extreme flows may not be as prevalent and whether they expect that there approach would need to be modified or not accordingly.

Thank you for this suggestion. We propose adding a discussion section about the transferability of model results in other regions. We have not tested this, of course, but it stands to reason that if hydrological regimes are vastly different, it is possible that the LSTM might need more data to be able to represent those more dissimilar catchments, as is the case in arid catchments. A very recent study (Nogueira Filho et al. 2022) has shown that LSTM and another neural network were able to perform better than conceptual models in a semi-arid region of Brazil. It is thus likely that the LSTM, given sufficient data and trained on data from similar conditions, would outperform hydrological models in those regions as well, although this hypothesis would need to be tested in future studies.

At this point, I am looking forward to reading the authors' point of view as I believe they have earned an opportunity to provide sound rebuttal comments as I feel the paper has the potential to be a valuable contribution to Hydrology and Earth System Sciences. Thus, for the time being, I would say that major revisions are necessary and required.

Please find additional suggestions/recommendations and editorial comments below that will need to be addressed thoroughly before the paper can be recommended for publication.

We thank Reviewer #1 for these comments and will now respond to the specific comments below.

Comments/suggestions/recommendations

P4 The following sentence, « In the Kratzert et al. (2018) study, the regional LSTM models performed on average just as well as the local LSTM with the median NSE difference of 0. » Local LSTM should be clarified compared to regional LSTM.

Good idea. This will be clarified as :

"In the Kratzert et al. (2018) study, the regional LSTM models (single models that can predict streamflow on a variety of catchments in a region) performed on average just as well as the local LSTM (trained specifically on a single catchment at a time) with a median NSE difference of 0."

P5 As illustrated in Figure 1 and Table 1, very large basins are included in the dataset, while including these basins during LSTM modeling has been quite a challenge since the input data are at the basin scale. How do the authors evaluate their results by assigning just one point to a basin with an average area of almost 31,900 km2?

The issue of scale is actually quite interesting and one we will detail more in the revised manuscript. For the "traditional" hydrological models, having a wide variety of catchments might actually be causing problems since they will be less similar to their smaller counterparts. It is probable that the hydrological models calibrated on the large catchments will not transfer very well to the smaller catchments due to differences in hydrological response. However, LSTM networks can make use of this information to build relationships using more diversity, allowing to detect patterns more clearly. Therefore, adding these larger catchments probably helps predicting flows on other catchments since it contributes extra data points in the model that will help avoid extrapolating during testing. As for the data being averaged at the catchment scale, this is indeed a limitation of lumped hydrological models, and feeding the same data to the LSTM seems to benefit the LSTM. One other limitation is that the LSTM applied in regionalization needs to have the same number of inputs as all the training sets, thus having variable numbers of inputs (e.g., weather stations) at each catchment would not be possible without using a much more complex LSTM structure.

P10 Why did the authors choose the leaky ReLU activation function? The authors should provide a table presenting the tested functions and values of the specificities of the LSTM model and the optimal ones; that would provide more insights to the readers.

Most parameters were adjusted using expert knowledge to focus on the hyperparameters that had a good likelihood of returning good results. Also, the model structure was generally made to be similar to that of Kratzert et al. (2018) given their excellent results. Therefore only a few hyperparameters were adjusted, as displayed in table 2. As for the LeakyReLU activation function, it was used to eliminate any possibility of generating impossible objective function values or exploding gradients. This was not common with ReLU but depending on the objective function choice and other hyperparameters, we did encounter cases when the model would not converge or would return undefined objective function values. LeakyReLU minimizes these errors, at the expense of a bit more computing time. This will be added to the text.

P11 Correct me if I am wrong, but according to the following sentence: « The twelve static descriptors presented in Table 1 allow the model to distinguish between each catchment ». Which one of them did the authors exactly use? Please provide another table introducing the list of twelve basin descriptors used for LSTM modeling.

This is an error related to the fact that in our first simulations we were using 12 descriptors, until we found that using more (many from recommendations in the literature) allowed for better results. We redid all the simulations but forgot to update this part of the text. It will be corrected in the next version. All 25 catchment descriptors listed in Table 1 were used in this study.

P11 According to the following sentence: « Static descriptors were normalized between 0 and 1 using a min-max scaler, while the dynamic variables were standardized by the mean and the standard deviation, which is a standard practice ». Did the authors include streamflow (target) during this normalization process? If not, how do they analyze their results after denormalization? Later, on the same page, it is mentioned, « The specific streamflow was used as the target variable by diving streamflow records by the drainage area, then converted from m3 s-1 to mm.d-1 .». Please further clarify.

No, the target streamflow data was not scaled using a min-max or standard scaler. Streamflow is not an input to the models; it is only used to evaluate the model performance (as for traditional hydrologic models). Only the inputs to the model were scaled in this manner. We will emphasize this in the manuscript. Working with input variables on a similar scale allows the model to converge faster with the use of larger learning rates. However, normalizing the target variable is not needed to accelerate the training and is typically not performed. On the other hand, streamflow is highly dependent of the drainage area, generating larger volumes for the same precipitation. While this information is available within the static descriptor, we found that including that knowledge upfront ends up accelerating the training process significantly instead of letting the model search for this correlation. The drainage area static variable remains useful for the model considering that hydrological processes with differ between small and large watersheds. Thus, streamflow is standardized by the drainage area, such that units are $m^3s^{-1}km^{-2}$. Then, by adjusting the length units, we can obtain the units mm3.s-1. This provides values that are hard to interpret/debug, so they are multiplied by the time units such that we obtain mm.d-1 units. This is what the LSTM model tries to reproduce and is trained on. Finally, once the LSTM returns a series of outputs for the pseudo-ungauged site, the reverse calculations are performed to obtain the flowrate to be compared to the observations. The fact that flows are compared on a mm/d basis means that the output can be tailored to any catchment.

P14 According to the following sentence: « This is important, considering that a strong performing hydrological model with the 3 | 4 best regionalization method is still outperformed on average by a relatively simple LSTM model. », the authors claim to use a simple LSTM model while using 2 LSTM layers each with 512 units, based on my experience, this is not considered a simple LSTM model. Please modify the text accordingly.

Thanks for this comment. The interpretation is correct, the text does mention that it is a “simple LSTM model”. However, this is relative, as in our opinion, an LSTM (even if the structure is internally quite complex) did not require much setting-up, calibrating, adjusting, etc. compared to other, more classical hydrological models. It is true that the LSTM model structure is quite complex compared to others, so this text will be modified to reflect this, i.e., that the LSTM model is complex but can be applied without a lot of work to represent the specific processes etc.

P14 Please be specific. According to the following sentence: « It is also important to note that the training (80%) and validation (20%) basins are categorized as such randomly, so the training step is performed on different catchments for each of the 5 runs #4a-#4e. », the authors should provide more details on how they couple this splitting approach with LOOCV, this needs to be clarified.

Indeed, this will be clarified along with the general comment #2 above regarding the training/validation/testing phases. Essentially, every time a model is trained, 1 catchment (pre-determined as the pseudo-ungauged basin) is removed from the lot. Then, remaining basins are split into 2 groups, i.e., training (80%) and validation (20%). This splitting is random in nature, thus when testing over a large series of basins (each of the 148 basins considered pseudo-ungauged one at a time) and while varying the model structure, it is clear that the stochastic component could play a role in the results. However, given the large number of such simulations, the expected variance from one test to the other should be very small.

P14 Figure 8 shows the sensitivity of the hyperparameter selection and the assessment of the LSTM model structure. The authors claimed that the performance generally increases with a more complex model structure, meanwhile Figure 8 shows that increments are very minor between the simple structure models and the complicated models. In real-world practices, training and calibration of complex models face major challenges, how do the authors explain the choice of the selected complex model?

In figure 8, we can see that the trend is monotonously increasing from model 1 to 7, in increasing complexity order. The median testing NSE increases from 0.740 for the simple model to 0.785 for the most complex model. Furthermore, each of the quantiles of the distributions are improving with each successive model. These types of improvements in regionalization are very significant. Therefore, the most complex model was selected since it outperformed the others, without requiring the modeller to integrate new physics/physical process representation. Simply by adding LSTM layers, the LSTM model was able to perform better in testing/regionalization mode, at the expense of computing time. It is true that building and training complex hydrological models, in the classical sense, requires a lot of effort. In this paper, we show that the performance can be better than what is obtained with hydrological models, but with little modelling effort from the hydrologist. Therefore, if the computing time is available, the more complex model is to be preferred. Especially since the training is only performed once for a given ungauged catchment application. A section to this effect will be added to the text, detailing this choice.

P15 According to the following sentence: « First, the nature of the LSTM model makes it extremely difficult or practically impossible to determine the logical flow of data between the observations and the predicted streamflow », readers may find it misleading since understanding the relationships between inputs and output of datadriven models can be achieved using sensitivity analysis. It is the authors' responsibility to provide such analysis as it would provide a way of following the logical flow of data. Thus, this sentence should be clarified accordingly.

Agreed, the sentence was not entirely clear. What was meant was that the physical representation of processes is lost in these deep learning models. How precipitation becomes streamflow is hard to track due to the numerous weights, non-linear functions and layers that add lags and biases at each step. Therefore, the best approach would be to evaluate the final trained weights and try to correlate them with expected hydrological variables, but this was not part of the scope of this study. Some papers have

already started showing these links (as stated in the manuscript) but it remains that following a precipitation value in an LSTM and seeing how it affects the streamflow for the next 3 days is very much convoluted compared to a classical hydrological model, where each process is explicitly defined. This will be rephrased in the next version of the manuscript.

P16 Based on the following sentence: « However, in this study, regularization failed to improve results ». Did the authors test all the possible values of dropout rates to reach such a conclusion? For instance, the value of 0.5 for the dropout rate has shown to be promising in improving the accuracy of streamflow modeling in other studies. Did the authors test this value?

Yes, dropout values of 0.1 to 0.7 were tested, and it was found that the values proposed here performed best (dropout of 0.3 for the LSTM layers and 0.1 for the dense layers). Dropouts simply drop some neurons during training to make the model more robust. On the other hand, regularization attempts to set some weights to 0 overall in the final model to remove noise from neurons that are weak (close to zero) to begin with, removing some influence from noise overfitting. This means stronger, more important weights remain, that are more robust to the signal. However, using regularization did not improve results in this study. In our case, perhaps the dropout rate was sufficient to provide this reliability during the training, or perhaps the fact that there were a lot of training samples meant that the model was able to converge on the signal without too much overfitting in the first place. This will also be added to the text for clarification.