Reviewer #2

General comments

This manuscript is a positive addition to the growing amount of research on the use of machine learning techniques in hydrological modelling with a focus on ungauged basins. The study compares an LSTM-based model trained over multiple catchments with three traditional hydrological models calibrated using several regionalization methods. Overall, the LSTM outperformed the traditional hydrological models at almost all catchments regardless of regionalization method used. This manuscript provides interesting results, is well structured, and was enjoyable to read. However, some additional clarifications throughout the manuscript would allow the reader to fully understand the chosen methodology and the presented results. Please see the specific comments below.

We would like to thank Reviewer #2 for their positive comments and for the suggestions on how to improve the manuscript. We have provided a point-by-point reply to all issues and comments below.

Specific comments

Introduction: As the authors rightly point out the LSTM has been used in several studies in recent years. However, the literature review is mainly focused on work conducted on catchments in North America and with limited acknowledgment of studies conducted in other regions (e.g., Choi et al., 2022; Nogueira Filho et al., 2022; Ayzel et al., 2021; Ayzel et al., 2020). Additionally, it would be beneficial to include a couple of lines near the beginning explaining that this study uses regionalization of hydrological model parameters specifically, and briefly defining what is meant by "hydrological model".

This was also highlighted by Reviewer #1, and we will indeed improve the literature review. We will expand it to a more global scale and will also clearly define our interpretation of "hydrological model" and regionalization. Citations that will be added include the following:

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.

Line 210-212: I am confused by the sentence “Each of these models was calibrated using the Covariance Matrix Adaptation Evolution Strategy (CMAES; Hansen et al., 2003) optimization algorithm in the Arsenault and Brissette (2014) study, and parameters are reused here to maintain the comparability to this study.” Was the HSAMI model not the only hydrological model used in the Arsenault and Brissette (2014)? Please clarify which parameters are reused, and how they relate to the calibration method and results described in lines 210-225.

This is a good catch. The models were actually used across a few papers instead of only the one mentioned previously:

Arsenault and Brissette 2014: HSAMI

Arsenault and Brissette 2015: HSAMI, HMETS

Poissant et al. 2017: GR4JCN

The information of the parameters that was taken from the other studies were the parameter boundaries for calibrations and not the calibrated parameters themselves since the models were applied to different catchments. This will be clarified in the text, referring to each study for each model and providing the clarifications regarding model calibration.

Line 262: Why was N=5 chosen (over other values between 4-8)? Please state the reasoning.

The value of N=5 was chosen purely due to it being recommended in the literature as a reasonable value in the 4-8 range. In many papers, values between 4-8 donors do not show any significant differences, and between 5-7 previous studies have shown that there is essentially no difference. So, N=5 was chosen to make sure the full effect of multi-donor averaging was at play while not using unnecessary computing resources to extend to 6, 7 or 8 donors. This will be clarified in the text.

Line 275: Please state how many catchments were classified as “poor” and thus removed when the filter was applied.

This will be added in the text. As seen in Figure 2, slightly more than half of the catchments have calibration NSE values below 0.7. This ranges from 84 to 89 basins depending on the hydrological model.

Line 307-308: “The twelve static descriptors presented in Table 1 allow the model to distinguish between each catchment.”. Highlighting these variables in Table 1 may make it easier to understand which 12 are used as input to the LSTM. Also land cover (%) is split into 7 entries in Table 1 but I think is only considered as 1 of the 12 static descriptors which is confusing.

Good catch, and we will copy a response given to Reviewer #1 to this effect for coherence and for your convenience:

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.

Also, the 7 land cover classes are all considered independently and count towards 7 of the 25 descriptors used.

Line 312: Please clearly define the training, validation, and testing catchments.

We have answered a similar question from Reviewer #1, and the response has been copied here for your convenience:

Indeed, this will be clarified along with the general comment #2 above (from Reviewer #1) 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. While the testing dataset could have been chosen as a fixed percentage (e.g., 20%) of all watersheds, using a leave-one-out-cross-validation (LOOCV) methodology was essential to compare results to previous studies.

Line 330: Why was model #7 chosen as the LSTM structure of choice? Please state the reasoning.

This point was also raised by Reviewer #1, and we have provided the following response, for your convenience:

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 0.74 for the simple model to 0.785 for the more 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. […] 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.

Line 374-375: Were non-linear relationships between catchment descriptors and NSE values considered?

At this stage, no, only linear relationships were considered, to see if there was a correlation (i.e., if perhaps larger basins reacted better than smaller basins, etc.). However, this was not the case, leading to believe that the LSTM was able to use these descriptors in a non-linear fashion to provide the good, basin-dependent regionalized streamflow.

Line 395-396: “relatively simple LSTM model”. Is this still referring to model #7 which is the most complex of the LSTM models tested? Please clarify. Also, on line 489 - “simple LSTM model”.

Thank you for this comment. Again, we refer to a response given to Reviewer #1 to this same question for your convenience:

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.

In the text, we will also clarify the meaning of “complexity” in the context of our study.

Lines 455-459: As discussed in the introduction (lines 119-126) traditional hydrological models and LSTM models show different behaviours in terms of performance for increasing lengths of data (e.g., the plateauing after 3 years of the GR4J model (line 122)). Please comment on the “fair-ness” of the comparison considering only catchments with at least 30 years of data are included?

This is a fair point, and we will add a discussion point to reflect on it. In theory, both he GR4JCN and the LSTM model have access to the same data and as such, the comparison is as fair as it can be. However, GR4JCN must compromise on the parameter sets to use to be “generally” good, whereas the LSTM has many more degrees of freedom to fit to various hydroclimatological situations. However, GR4JCN has a predefined structure where processes are directly defined, whereas the LSTM must build its internal structure using its more numerous “parameters”. We will add a discussion point detailing the fact that not only the LSTM is to be favored due to the long time series of available data, but also that it can ingest data from many more catchments as well, whereas GR4J is limited to containing information from one catchment at a time.

Technical corrections

These technical corrections will also all be addressed in the revised version of the manuscript:

Line 12: Suggest changing “A series of …” to “a set of …” as series implies that there is a sequential element to the methods.

Line 12: “regionalization methods are applied”

Line 180-181: “Environment and Climate Change Canada (ECCC), and the United States Geological Survey (USGS).”

Line 232: Suggest changing “for each scenario” to “for each of the 18 scenarios” for clarity.

Line 288: “have difficulty remembering”

Line 315: “then converted from m.s-1 to mm.d-1” (as the division by drainage area would already have removed two spatial dimensions).

References

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.

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.

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.