We would first like to thank the Reviewers and Editor once again for this second review of our manuscript, which will again help improve this paper. As in the previous round, this document presents the modifications made to the original manuscript in response to the reviewer comments and suggestions, with the original comments in black font and author responses in blue.

Reviewer #1

Thank you for the recommendation and highlighting this typo, which has been fixed in the revised manuscript:

Line 338: "and converting to mm.d-1"

Reviewer #2

General comments

The authors provided satisfactory responses to most of my questions. They clarified the utilization of LOOCV, and the training of the LSTM with the given dataset. The upgraded literature review is well organized, yet they could have surveyed so many other relevant references (from North America) addressing the same topic and not solely focusing on those studies conducted by one specific research group. Thus, to further improve the quality of the manuscript, I believe, there are still a few issues that need be dealt with before the paper is deemed acceptable. Thus, for the time being, I would say moderate revisions are necessary and required.

We would like to thank Reviewer #2 for their comments and suggestions for improving the paper even further. We have provided a point-by-point reply to all issues and comments below.

Specific comments

P4: On this page, the authors attempted to complete the literature review, the work done by Feng D, Lawson K, Shen C. (Prediction in ungauged regions with sparse flow duration curves and input selection ensemble modeling. arXiv preprint arXiv:2011.13380. 2020 Nov 26) should be cited since it is related to the concept brought in this study and has been applied in North America as well.

Indeed, for the literature review we stayed mostly with studies looking into continuous streamflow prediction rather than regionalization of hydrological indices, prediction (in the forecasting sense) and other such models. For this specific domain, we have not found any more pertinent studies. The reference recommended was however added in lines 98-99:

"Feng et al. (2020) showed that using regional flow duration curves as predictors in an LSTM model improved the prediction in ungauged basins skill over 671 CAMELS basins compared to an LSTM model without the flow duration curve inputs"

However, it is a preprint since 2020 and the paper refers to supporting information that does not seem to be available in the public domain. Therefore, we are unsure this reference will remain after the editing/typesetting stage. This is why it was not added at a prior stage.

P11 Line 335: according to this sentence written by authors: « Streamflow was not scaled itself using this approach, since the model outputs and target values are not part of the model training computations and thus have no impact on the numerical convergence efficiency. ». This sentence raises two questions: (i) If the authors have not used the target values during the training, how do they train the model? Indeed, one of the essential elements to train and calibrate ML models is the target since the model cannot decipher the physics. (ii) I believe, the rationale behind the authors' work; that is the normalization of the target is not necessary, is not quite correct. Indeed, a target variable with a large spread of values, in return, may result in a large error gradient values; causing the weight values to change significantly, leading to an unstable learning process and the occurrence of a rather slow learning process. Please correct the text accordingly to avoid any confusion.

Sorry for not clarifying sufficiently. First, yes, streamflow is used for training. This is indeed the entire basis of the training and very much required to obtain the model hyperparameters used for regionalization to the ungauged basins. This has been clarified in the text.

Second, the normalization aspect can be a problem in some cases when the target variable spans multiple orders of magnitude and can cause convergence issues. This was not a problem (and thus not required) in this study for a few reasons. First, the LSTM model was trained using scaled streamflow (streamflow divided by catchment area). This means that the range is already quite reduced and well outside the range of magnitudes that would cause problems in the gradient estimation / exploding gradients. Second, the learning rate was dynamically adjusted to progressively diminish every 5 epochs such that the convergence was controlled the entire time. Therefore, the normalization of streamflow was not required in this study.

This was clarified in the text as follows in lines 342-348:
"The target variable of streamflow was not itself scaled using this approach, since the model output and target values are not part of the model training computations and thus have no impact on the obtained results. Instead, the specific streamflow was used as the target variable by dividing streamflow records by the drainage area and converting to mm.d-1. This was done to allow combining information from the multiple training catchments during the LSTM training since all streamflow values were now represented in an area-independent depth unit, while at the same time ensuring all values had similar magnitudes to avoid convergence problems."

And the learning rate was also detailed in lines 365-366:
"In all cases, a decaying learning rate was implemented to ensure proper convergence of the training algorithm, refining the learning rate as a function of the number of epochs."

P12 The response to the question about using Leaky ReLU was satisfactory. Yet, there is a need for additional clarification. If the LSTM structure used in this study is like the one proposed by Kratzert et al. (2018) (It appears that the authors just used two LSTM units to make a more complex one), the authors must mention the work done by Kratzert et al. when they are discussing the model structure. This way, more details will be provided to the readers to extract the information regarding model structure (e.g., hyperparameters).

Thanks for this suggestion. We have added information regarding the Kratzert et al. (2019) setup in lines 323-328:

"In their paper, Kratzert et al. (2019b) tested multiple model structures and hyperparameters, including up to 256 units per LSTM layer, both for one and two layers, and with dropout rates ranging from 0.0 to 0.5 and input sequence lengths of 90 to 365 days. They finally settled for the model that provided the highest median, which was a single-layer, 256-unit LSTM with a dropout rate of 0.4 and an input sequence length of 270 days. However, the static descriptors were directly embedded in the LSTM layers, as opposed to their addition in a separate, parallel branch that is also tuned during training in this study."

P16 According to the following sentences, « For example, in this study one catchment has a much larger area than almost all the others. For a hydrological model-based regionalization approach, this might skew the regressions between catchment descriptors and model parameters. LSTM, on the other hand, are strongly non-linear and are thus not bound to these limitations. They could also use these data to better predict streamflow processes at scales between the small and large catchments. ». Is this where the authors explain the use of large catchments in LSTM modeling? Since I am not convinced with the explanations provided by the authors on how they rationalize including large catchments in their modeling. If that is because of the complexity of the model, it should be clarified thoroughly. Please provide more reasoning.

Yes indeed, this is where we justify the added value of large catchments in the dataset, and we have clarified further the text as follows, in lines 484-488:

"This is because neural networks in general, including LSTM-based neural networks, are particularly good for interpolating within the domain they are trained to represent but can be unpredictable while extrapolating outside of the parameters of their training dataset. Therefore, adding catchments with a wide array of properties confers the ability to establish relationships that other methods simply cannot attain by widening the domain on which the model can interpolate."

P17 In the part highlighted in blue, can the authors verify what do they want to explain to the readers? Since it is not quite clear which point (within the comments) is addressed here.

This point was in response to a previous comment (comment for lines 455-459 of Reviewer #2 in the previous review round) regarding the length of available data being a factor in the obtained results. It would be expected that a classical hydrological model would fare better than an LSTM if only 2 or 3 years of data were available, for example, because the hydrological model has a priori knowledge of the expected physics while the LSTM would need to build that model itself from the data. However, a hydrological model can use only the data from the catchment itself, whereas the LSTM can learn from other catchments as well, meaning it can also perhaps work well on a series of catchments that have few years of data each, if there are enough such catchments.

We have clarified this in the text in lines 510-520 of the revised manuscript.

P28. To avoid any confusion, please correct the information provided in the caption of Figure 3. It is mentioned that N-1 catchments are used during training which is not correct. These catchments are used during training and validating using an 80-to-20 split, respectively.

Thanks for highlighting this discrepancy. It has been corrected in the text as follows:

"Performance of hydrological models calibrated at each of the 148 study basins individually against the performance of the LSTM model in leave-one-out cross-validation (LOOCV) where the ungauged basin in question is not included in the set of basins used to train the LSTM and where the LSTM is trained and validated on 80% and 20% of the gauged basins, respectively."