I would like to respond briefly to the author's responses before this goes back to the editor. I think this is important to state up front that these responses are not sufficient to warrant publication of this paper (in my opinion). The authors' responses indicate a very deep misunderstanding of the models that they are working with, and also a deep misunderstanding of relatively simple and basic ML procedures. I sincerely hope that they will take the advice I gave in the original review, because this paper is not close to being publishable as it stands, and the authors responses only further indicate that we are not approaching a reasonable experiment.

(author comments in purple)

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.

This is because you did not do hyperparameter tuning correctly. You did not tune most of the important parameters, and you did not do a full sweep. Indeed, hypertuning an LSTM is extremely important and will change the results significantly. It is clear from what you reported that you did not do this correctly, however the fact that you did not find it useful is by itself an indication that you did not do it correctly. Hypertuning is an extremely basic part of any ML workflow, and the fact that such a basic thing was done so poorly is a hint that there may be other serious issues with this study (indeed there are many).

Hypertuning must be done with a full sweep of parameters on a large dataset separately for each model. This is non-negotiable. If this is not done then the results of the study are meaningless. Hypertuning matters – it is a problem that the authors are seeing/arguing otherwise.

We would assume that the reviewer's term "LSTM's behavior" could be translated to "LSTM's performance measured by the median KGE"

No, this is not correct. What I mean is about how the activation functions in each gate "learn" to span the range of possible inputs and responses. I suspect from reading this response that the authors have not spent much, if any, time understanding the behavior of the gate structures in their model. Have the authors plotted or visualized their activation functions? Or their forget, input, and output gates? Have they spent any time looking at the range and independence of their cell and hidden states? I suspect not. Please understand that several years worth of this kind of effort is underneath all of the studies published by Kratzert et al.

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.

No, again this is not a correct interpretation of my comment, and it (again) belies a deep lack of understanding about basic ML workflows. You should hypertune on a sample that is as

representative as possible of the use case. In the case of an LSTM for streamflow, you will use large, diverse datasets to train the model, so hyperparameter tuning will be done on similarly large, diverse datasets. For local models, you will hypertune each individual local model separately on held-out data from that catchment.

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.

Yes, this is because you've done hyperparameter tuning incorrectly. This is a flaw in your workflow that needs to be fixed before the results can be meaningful or trusted. Please see other studies (preferably by ML experts, not by other hydrologists) on how to do correct hypertuning with the LSTM for streamflow. Kratzert uses three data periods for each catchment. There are alternative ways to do this, but the hyperparameter tuning must be done in a meaningful way, and it must be done separately for each model that is benchmarked/compared. None of that is done correctly in this study.

About cell state size, you said: *"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."*

First, why are you trusting a secondary source over a primary source? Kratzert and Hochreiter are ML experts and Lees is not. Kratzert and Hochreiter spent years developing, training, and tuning the model, and Lees did not. Lees did a good job, but his paper is just an application study. Kratzert has spent years building and running operational versions of this model at local, regional, and global scales for many different purposes (operational models, research models, private sector, government, etc.), as well as exploring the model in depth in a series of publications. Kratzert worked directly with the world-leading LSTM expert. Lees' paper is a one-off application study. I'm not saying that Lees' study is wrong or flawed (it is a good study), but I wonder why you would choose that one in particular to model your experimental design from. In response to your comment about publication date, Kratzert has many papers newer or contemporary with Lees. Leaving aside the question of how you choose which references to model your study after, in the present case you have no idea what cell state size to use because you didn't test this parameter. No matter how you slice it, it is inappropriate to not hypertune one of the most important parameters in the model.

"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."

This is not true. This sounds like things we very often hear from hydrologists who do not have very much or any experience using these deep learning models. There is very little overhead for using a regional model that does not also exist for using local models, and an extremely large amount of overhead for using local models (fine tuning is cheaper than hypertuning). Training regional models is not expensive, and can be done on a laptop (unless you design your code

inefficiently, however you should be using the NeuralHydrology github repository unless you are an expert at building ML models – there is a huge amount of work underneath setting one of these models up correctly, and it is easy to do it quick and wrong).

I response to this comment, if/when the authors actually become familiar with these DL models and start using them regularly, you will find three things:
(1) There is no significant difference in the "size" of regional vs. local models (the difference in size is a few MB, and the increase in the number of free parameters does not cause performance or behavior issues).
(2) It is not harder to use regional models than local ones (in fact, the opposite because fine tuning is cheaper than hypertuning).
(3) Regional models are usually more accurate (when done correctly) because any regional model can be tuned as necessary.

"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."

This is irrelevant, because there is enough public streamflow data to train a base model. There is no reason to use "national" data because you can just tune your base model. Use the largest dataset possible. This, again, is an imagined constraint that we hear often from hydrologists who do not have a strong understanding of how these models work. I'm going to be a little personal here and say that it is frustrating to respond to these types of comments, when it is extremely clear that the authors don't actually know this from experience (I'm so sorry to be direct, but when the authors gain experience, they will see that this statement is incorrect).

Train your base model on a large-sample datasets, and then do whatever you want (i.e., tuning, etc.) with those base models using whatever dataset you are actually interested in. This is how these models work best. We have a lot of experience doing this (including operational modeling at all scales from single-catchment to global).

"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"

This is, again, not really correct. I want to emphasize that there is no well-established way to deal with dams and reservoirs in (any type of) hydrology model, including deep learning (this is a problem that we are currently working on). The problem is that the authors did not make this comment from any actual experience using LSTMs. I know this not only because their use of LSTMs throughout this paper, and as indicated in their replies, is in many places contrary to (well-known) best practices, however, I also know that the authors are not making this claim from experience because I've spent months working on this (modeling dams with deep learning), and locally trained models are one of the worst options that we have tried.

Again, I'm so sorry to be personal about this, but I am really frustrated at having to respond to comments made by hydrologists with no actual, meaningful experience with these issues. I am

asking (begging) hydrologists to stop imagining things, and make their decisions about which and how to use different models based on evidence and experience. I'm so sorry to be direct, but I'm asking the editor and authors to please understand how pervasive and pernicious this problem is in the deep learning hydrology community (pretending to understand the limitations of different models without actually exploring those limitations in depth). I would welcome a published study comparing local vs. regional models for managed catchments, but such a study will need to be done using best-practices (including rigorous hypertuning, rigorous benchmarking, and rigorous fine tuning). Such a study could (possibly) be interesting even if it *didn't* include any of the more sophisticated ways to approach the problem (e.g., reinforcement learning, transferable/tunable head layers, multi-output learning, data assimilation, etc.)

For what it's worth, we use LSTMs for managed catchments operationally at both the global and local scale in both the public and private sectors. This is after years of experimenting and practice. I'm not saying that I know the solution, but I am saying that it is transparently clear that the authors have made this claim without actually trying it in a serious way.

"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"

Definitely, if you think you have something new please test it. I'm not seeing anything in this study that represents a new idea in terms of how to train models (we've performed and published rigorous local vs. regional LSTM studies multiple times in the past, using what appears to be an identical training procedure except that we did real hypertuning). I don't want to discourage you from testing whatever ideas you might have, but please do the experiments in a way that allows us to be sure that the effects we are seeing aren't due to taking shortcuts. If you are computationally limited (e.g., can't do full hypertuning for each model that you want to test) then this means that you can't do the experiment - full stop. Please don't publish half-done studies, it clouds the literature with misleading results. I want to be very clear: I am happy to see the authors testing these two hypotheses outlined in this paper, but they need to use an experimental design that will actually answer the questions that they want to ask.