Dear authors,
I read your manuscript with great interest and think it is overall clearly structured and well written, congratulations on that. I do have a few comments, which I sincerely hope that they enhance your manuscript, none of which is anything serious. I also left a couple of minor comments below. Some of these are probably enough to be answered in the reply to my review and don't need a change in the manuscript.
The only thing that I'm not 100% sure about, and which I think should be the decision of the editor, is the following: Most of the manuscript reads to me as a technical report of your experience/participation in the streamflow forecast rodeo, which I think is a wonderful thing to publish. I'm just not sure if this manuscript should be published as a technical paper therefore, or, as is, as a research article. I do not have strong feelings about this but I thought I would mention it.

I will structure my review as follows:
- First, I try to list some general points with which I struggled during my review and which I hope can guide the authors to enhance their manuscript.
- Second, I list a couple of line-by-line comments. Some of which might need a change in the manuscript, for others it might be enough to discuss this here in the review.
Looking forward to an interesting discussion,
Frederik Kratzert
Thank you for taking the time to review our manuscript and for providing us with some very helpful suggestions, from which we believe the manuscript will benefit greatly. We have responded to these points individually below, in red.

General points:
- Data and model setup: I struggled a bit to follow the exact model setup and which data, in which temporal frequency, is used in the LSTM. Some of the information is already there, but distributed across the manuscript. I think it could increase the readability to have a dedicated paragraph (e.g. in Sect. 4.3, where some of the info is already available) that simply states: "We use A, B, ..., C as model input at a temporal resolution of D. The model is trained to predict E days (hours?!) of streamflow forecast, using F features during the forecast period.". Again, some of the information is already there, but distributed across different sections (e.g. Data and Model sections). I found myself multiple times searching back-and-forth in the manuscript, to try to understand exactly what you do. E.g. in line 155 you say that streamflow was available in 3-hourly resolution, L. 248f states that all (gridded) data was averaged over the catchment "at a six-hourly frequency". From looking at the table on page 11 and L 249 of the manuscript, I can see that your input data frequency was 6 hourly, but the same table also says that you have a single model output. How exactly are you generating the forecasts for all *n* forecast timesteps?
We agree that the manuscript could really benefit from a summary of the overall setup and operational deployment, which we will include, as suggested, in the introduction. We will also act on similar comments by the other two reviewers to clarify our methods throughout, for example including a table of input hydrometeorological variables.

- As one of the authors of some of the LSTM literature that you cite, I feel like I need to comment on your general training setup. It seems like your setup was mainly guided by our first LSTM paper (2018), rather than any of the more recent papers (which you cite as well). You are probably aware that many things have changed since then and that your modeling setup does not really represent the best practices when working with LSTMs. For example, many studies (not only our papers) have shown that LSTMs really excel, when being trained on data from multiple basins at once and not on a gauge-by-gauge level as done in this manuscript. Even if the modeler is only interested in one (or a few) basin(s), it is better (with respect to the performance in the basin(s) of interest) to train on a larger set of basins. I think if you want to avoid running additional experiments, it would make sense to include at least a discussion on this topic.
We agree with this comment. When I set up the LSTM originally a few years ago, the methodology followed your 2018 paper (which was then more-or-less the only practice, let alone the best practice). Of course, we do appreciate that accepted best practice has changed in the meantime and will gladly discuss this in more detail in the revised introduction. Given that the per-basin LSTM results are still reasonably good, we hope that the reviewer won't insist on a complete multi-basin rerun of our entire project!

- Similarly, and because you mention the computational expenses of training LSTMS, a comment on your model architecture: I know I used stacked LSTMs myself in the past (in the 2018 paper) but I was young and inexperienced. Since then, I spent many hours comparing different architectures and I never found a setup where it was worth using more than a single LSTM layer (which is what we have done since then in all of our publications). This is not critical and I don't want you to rerun experiments with different architectures, but it might be worth thinking about that when commenting on the computational expense. My experience from multiple large-sample model intercomparison studies is that the LSTM is by far the fastest model to train, compared to optimizing any hydrology model on scale. And the speed increases drastically if you use a single LSTM layer vs 3-stacked layers.
This is a very good point; one which we will gladly discuss in the revised manuscript. Regarding the benchmarking of (a) stacked vs single-layer LSTMs and (b) single-layer LSTMs vs hydrological models, are there particular references you would recommend? I was hoping to see some relevant data in your 2019 paper (https://hess.copernicus.org/preprints/hess-2019-368/hess-2019-368.pdf) but it didn't seem to touch on the issue of training speed.

Line-by-line comments:
- L. 43 "Schmidhuber et al. 1997" missing in the Reference section and "Hochreiter and Schmidhuber 1997" is wrongly formatted in the Reference section. Usually only "Hochreiter and Schmidhuber 1997" and "Gers et al. 2000" are cited as a reference for LSTMs.
Thanks – we will make these changes.

- L 48 "Chung et al. 2014" No need to cite this paper. This problem was the reason the

LSTM was invented and is already discussed in “Hochreiter and Schmidhuber 1997”, which would be a more appropriate reference here.
OK – this is a reasonable substitution that we’re happy to make.

- L 57 “..even if the models were trained on multiple basins…” This is linked to my comment above. It is not “*even*” but “*because*” the models were trained on multiple basins. There are multiple studies (some of which you already cite) that show that the LSTM is able to transfer (learned) knowledge across basins. From my experience, LSTMs trained on a per-basin level are often actually not that much different from hydrology models. Another related reference that shows how model performance of LSTMs increase with the number of basins is the “The proper care and feeding of CAMELS: How limited training data affects streamflow prediction” by Gauch, Mai and Lin (2021).
We will include these changes in our expanded discussion of the literature (see also response to major point 3). Thank you for the reference.

- L 59 “Extending this work, Gauch et al. (2021)…” Not really sure if I understand this sentence correctly. In Gauch et al. (2021), we extended the work from the other papers to models that are able to predict streamflow (or anything actually) on any given temporal resolution (and at multiple resolutions at the same time). What do you mean with “used a reanalysis framework to demonstrate the predictive power of LSTMs in streamflow modeling”?
We added this to demonstrate that work had been done on predicting streamflow using reanalyses, and not just observations (i.e. work deriving from the CAMELS datasets). The summary suggested in this comment is helpful though, so we will include it in our revision.

- L 149 “as close as observation as possible” -> “as close to observations as possible”. But is this statement even true? Does GloFAS produce simulations that are “as close as possible” to observations? You mean as close as possible “for GloFAS” (i.e. GloFAS can’t produce any better simulations) or as close as possible as “any model” can get? I would question the latter.
We definitely mean “for GloFAS” here, and will rephrase this in the revised manuscript.

- L 174 I shrugged when I read this passage. Personally, I really have problems to see that anything better than the long term mean is “skilful” and this is also not what Knoben et al (2019) said. Knoben et al. (2019) say that this threshold is often used to differentiate between “good” and “bad” models but they follow this with an explanation why such a global threshold is problematic in e.g. different kinds of flow regimes. Maybe you can at least extend the discussion around these thresholds a bit.
Yes, of course any definition of “skilful” is typically quite arbitrary. I certainly agree that just doing better than the climatology/mean doesn’t really indicate meaningful skill, a definition that has irritated me in seasonal model evaluations for years. So perhaps this usage is a case of Stockholm Syndrome, but I’m happy to discuss it a bit more – including correcting the statement attributed to Knoben et al (2019) – in the revised manuscript.

- L 204ff Great explanation!
Thank you!

- L 219 Eq. (7), Can you provide some explanation why you used a combination of NSE and KGE as your optimization function.
This is a good question, for which we will include an answer in the revised manuscript. Essentially, using NSE alone leaves the optimization procedure vulnerable to incorrect local minima (e.g. incorrect mean). However, we also found that using KGE alone tended to result in a bias correction that weighted correct mean and variance too highly compared to correlation (not useful for improving forecasts). We found that combining the two improved the weighting more in favour of correlation, while avoiding spurious local minima.

- L 226 Eq(8) Not entirely sure if I understand what I see. Is it correct that this matrix suggests that rather than including neighboring pixels (i.e. those in direct proximity to the gauge of interest) it relies on some distant pixels in all directions of the gauge? Are some of these pixels "downstream" of the gauge?
Yes – this interpretation is correct. All pixels are in the close neighbourhood of the gauge, but some are indeed downstream. This might seem a bit of a strange choice, but there are two advantages: (1) it is quick to set up, no additional topology data or preprocessing is required; and (2) we think (although we have not tested it) that it could help reduce temporal biases in some cases, for example, where the model puts the flow "too early", downstream information can actually be useful in correcting the bias. Where downstream information is not useful – as in the example given for NFSW4 in the paper – the weights collapse to zero (or very close) anyway.

- L 235 (equation) Is this based on (personal) experience or was this specifically tested for this study? I think it could be helpful to explain this with one sentence.
This was specifically tested at the beginning of the operational phase when we realised that the differences between GloFAS and GloFAS-ERA5 were leading to an overenthusiastic bias correction. We will mention this in the revised methods section.

- Figure 2: Great Figure and interesting results!
Thank you!

- L 262 "Note: snowc not available in IFS output" Is this an artifact from the manuscript preparation or otherwise can you extend what you want to say here?
This is left over from the original manuscript preparation and will be removed in the revision.

- Sect 5.1.X "Evaluation of the test period": I might be missing something but can you explain what exactly is shown as a simulated hydrograph? The model is built to provide a 10-day forecast right? So for every calendar day, you would have multiple forecasted values. Which value do you pick for these plots?
All models are driven with ERA5 during the testing period, since they are not being run

in forecast mode and we simply want to test their respective abilities to replicate observed streamflow. This contrasts with Figs 6/7/8, where the models *are* run in forecast mode (driven by IFS output), and for which we plot hydrographs at three different lead times.

- Fig 6, 7, 8 are a bit hard to read because of the small figure size and the many overlapping lines.
We appreciate that these figures are quite compact – but they are vector images so readers can get finer detail by zooming in. If the reviewer thinks that removing a timeseries from each (e.g. 2-day lead time forecasts) would be helpful, we could do that.

- L 312 I think you want to remove the "(CHECK)".
Thank you for spotting this – will remove.

- L 420 "training them [LSTMs] is computationally very expensive". My understanding/experience from participating in various large-sample benchmarking studies is that training LSTMs is much faster than optimizing hydrology models on a similar scale. This is probably related to your specific LSTM architecture (3-stacked layer), which by itself might not be necessary.
This comment is very similar to the last of your general comments. Again, we will happily include this caveat in our revised discussion.