Dear authors, I am thankful for your responses and I want to say that I am sorry if I appear to be the “mean reviewer 2” in this revision process.

Let me start with an TLDR;

- I think we can stop arguing here and conclude that we need to agree to disagree. I think I could go on forever but it probably adds little to this review process. So to save us all time, I will stop arguing after this review and I only reply here to clarify a few of the statements/misunderstandings.
- That being said, I think the first point, regarding data splitting (TLDR; use the results of the test splits in the paper, not the averaged results of training periods) should be taken seriously.
- One last thing: I mentioned this to Shijie at the EGU in personal communication and I don’t want this review process to negatively impact what I said: I think this is an exciting study and I am happy to see people using LSTMs for these kinds of studies. I think it would have been easy to make this paper even more exciting but I understand that I need to wait for another publication from your group to get answers to my questions.

Data splitting

I am very happy to hear that running your analysis only on the test splits did barely change the results. What I don’t understand is, why you did not change the manuscript so that these results are used in the paper. I don’t think that I need to search for literature here to say that in any modeling study, you should use independent test data (which you have!) to analyze/interpret your model. We already know that this is not affecting anything in your results, so it should be a no-brainer to change this and standard practices in statistical analysis with models.

Input selection

I think it is somewhat funny that on this point you argue with NSE performance (that it did not or did negatively affect the NSE, when adding more inputs), while in the other points (e.g. multi-basin model) you say that performance is not really important.
Anyway, there is not much to add from my side, I think it would have been a great addition to the paper to include and analyze more input features, especially because it seemed like you already did the experiments. You yourself said in the first rebuttal

“During our preliminary tests, we had run models with daily averaged sea level pressure, relative humidity, and radiation as additional inputs, which did indeed lead to more clusters in terms of feature importance patterns”,

which to me sounded exciting and I was curious to hear more about clusters that are different from previous literature and about interpretations/hypotheses of these clusters. You decide to keep it simple, your choice. If I understand you correctly you want to come back to this in a future publication and I am looking forward to diving into the results then.

Model setup

Let us just say that we disagree. I certainly do believe that, no-matter what you want to do, if you use a tool, you should use that tool correctly. Following your argument (which tries to say the other extreme is possible) I could make such a study with uncalibrated hydrology models and then say that my results have any meaning?!

From a very basic point of view: I hope we can agree that a model that is better in modeling the task at hand (meteorological forcings in, discharge out) should have a better understanding of the underlying system/processes, right? Now if I want to make a study that investigates the underlying process understanding of a model, then why would I willingly pick a model that has a worse understanding of the underlying processes? If the model would not have a worse process understanding, it would not be generally worse in predicting discharge, or would it? Since your study is about model interpretation of process understanding, I do not see how this is decoupled from performance. But I see how this example will be used again to say “he only cares about performance”, which is wrong. But I’m not so naive to think that good process understanding is decoupled from good performance.
Your line of argument is the ease of interpretation, I got it. But what I am saying is that this is a bad argument. And again, this is not “because the regional model has a higher NSE”, but because, without any doubt, this model has a better process understanding. And your study is about analyzing such a process (flood generation). Funnily enough, the one thing where single-basin models have their biggest bias (compared to regional models) is flood peak prediction, which I also mentioned in my previous review (see point on saturation). One could argue that having a (negatively affecting bias) in flood predictions is also affecting the flood generating processes in the model.