Comments by MH in *blue*, our responses in black.

Summary

The Opinion paper addresses a common issue and misconception in the application of LSTM models for hydrologic streamflow prediction: Often, LSTM models are trained and evaluated on only a few basins or even a single one. This leads to sub-optimal performance as LSTM models benefit from being trained on a large variety of data - as they are increasingly available in hydrology. Therefore, the authors suggest to conduct LSTM training with such large-sample datasets as best-practice (outlining that additional fine-tuning might be an option for single or small set basin applications). Further, the paper focus on training data diversity and optimal setup-up of training sets.

Evaluation and Recommendation

The Opinion paper covers a timely topic since LSTM models are state-of-the-art tools for streamflow prediction and a variety of other tasks in the broader geosciences. With LSTMs being increasingly used (which is also shown in the paper), the presented topic is important and meets the community's interest.

The manuscript is well written and referenced. The codes that were used are freely available and data sources are referenced. The figures are of good quality. Yet, the current manuscript requires some specifics to be addressed (see below). In particular, section "5 Is hydrological diversity always an asset?" addresses a very important topic – at the same time, it would benefit from some iteration as is also specified in the comments below.

I recommend publication after minor revisions.

Specific comments

> *l.18-19: "We do not mean top-down vs. bottom-up in the sense discussed by Hrachowitz and Clark (2017)." Please specify briefly their definition of top-down and bottom-up.*

Response: Thank you for this suggestion. We will do exactly what the reviewer suggests in the revision.

> *l.28-29: "We see no reason why…" Please briefly elaborate on these reasons.*

Response: Thank you again for helping us to add clarity to the writing (sincerely). We will do exactly what the reviewer suggests in the revision.

> *l. 32-34: Please rewrite for clarity, e.g. split up in two sentences.*

Response: We will split the run-on sentence and do what we can to make the passage easier to understand.

> *l.50 ff & Figure 2: It is stated that 400+ catchments from CAMELS were modelled with mHM and VIC for comparison. How many did overlap with your 531 basins? There has to be a large overlap anyway with 671 basins in CAMELS in total, but I think it would be an interesting information to know. Or better: why not showing the cumulative plots between only the basins that are in both the VIC+mHM set and the LSTM set? This would make the comparison stricter.*

Response: This is an artifact of some legacy benchmarking experiments where some of the physically-based models only ran over a subset of the CAMELS basins. The 400 is a strict subset of the 531. Our reasoning in not making the basin groups match exactly is that we wanted all of our ML-based CAMELS results to be identical throughout the paper. The "fair" comparison in this figure is Figure 4 and Table 3 in Kratzert et al. (2019b). Changing the basin group in this figure does not change the message, but we agree with the reviewer that it is probably cleaner to make these subplots match. We will do that in the revision.

> *l.63: "to 1 (-1)" -> unclear, is this [1 -1]? Please specify.*

Response: Yes, it should be (-1, 1). We will fix this notation in the revision.

> *l.64-65: "size equal to the number of cell states." -> Please add one or two explanatory sentences. This refers to the model architecture and might not be clear to all readers.*

Response: Thank you again for offering suggestions about how we can make the manuscript more accessible to more readers. We will take the suggestion in the revisions.

> *l.73: "no model captures all of the extremes" -> Agreed. Yet, I wonder whether the chosen basin in Fig. 4 represents "extreme events" well. The hydrograph shows a rather regular pattern with a peak flow ranging roughly between 5 and 15 mm/d every year. I think a more irregular hydrograph with, e.g., one or two intense peaks (or missing peaks in some years) would better illustrate extreme events.*

Response: It is a good suggestion to add more examples to this figure and discussion, to help illustrate diversity. We will do that in the revision.

> *l.83-84: "In other words, even these 531 basins are most likely not enough to train optimal LSTM models for streamflow." Why do you think that is? Please elaborate. Are there indications for an upper limit of the objective value and a corresponding sufficient training size from research with CARAVAN that could be shown here? (This last point could also be part of conclusions and outlook)*

Response: The reason is because we have not found the saturation point in terms of improving skill scores from training on more basins. We could use Caravan, however there is currently no "official" LSTM benchmark, and we want to make sure that the message in this paper is as clear as possible. We also do not want to run the risk of readers thinking that the results might be different because we used a new dataset. We don't think this is the right venue for showing LSTM results with Caravan.

> *l.85ff: This section covers quite a range of aspects. I suggest to restructure it a little, e.g. into subsections: One that contains Figure 6 and the corresponding text, and one that covers conjectures about effects of larger datasets and data split approaches. Nonetheless, these are interesting points to be discussed since they also might pave the way for further research.*

Response: We will look for a way to subdivide this section to make the message more concise and focused.

> *l.87ff: "more is always better, as far as we have seen), and variety refers to the (hydrologic) diversity of data." -> Both points, volume and variety, are important to be pointed out. There is an issue that I think could be mentioned and discussed in this context as well: class imbalance (even if this is not a classification problem). This might also be part of the explanation behind the things discussed in the rest of section 5. In the CARAVAN-paper (Kratzert et al., 2023), there is a histogram showing the distribution of catchments over the different climatic zones of the earth and the corresponding distribution in the dataset. There, a class imbalance is visible which indicates why predictions of certain climatic basin classes are better or worse since they are overrepresented or underrepresented.*

Response: First, there are no model results reported in the 2023 Caravan paper (and we have not published any modeling results on Caravan anywhere else) so we are a little unclear on the connection that the review is making between model performance and class imbalance in the Caravan dataset. That being said, we are not convinced that the reviewer is correct that class imbalance is an issue here. We have not seen any empirical evidence that suggests that this is an issue.

> *l.148: "training (1 October 1999 through 30 September 2000), validation" -> mix-up of dates? This is a very short training period.*

Response: Thank you, this is a typo.

> *l.210-211: "without carbonate rocks fraction and the seasonality of precipitation" -> I agree on dropping the seasonality of precipitation. But I would assume that carbonate rock fraction and related karst flow properties might be an important feature to be included in the clusters. Did you investigate in which clusters those*

basins fall, that have a carbonate rock fraction value larger 0 or above a certain threshold? Maybe having a dedicated "karst" cluster might be an option?

Response: Thank you, this is a typo.

l.215: "into 5 groups" -> should be "into 6 groups", right?

Response: Thank you, this is a typo.

Fig A4: Interesting to see how geographically aligned the different clusters are. (Apart from separated small accumulations apart from the bulk of a certain cluster – like the small groups of cluster 2 (orange) and 6 (brown) north from cluster 1 (blue)). With 6 being the detected optimum number of clusters, did you look into the neighboring 5 and 7 clusters with respect to the spatial distribution on the map and performance gain/loss of the model?

Response: Yes, the attribute clusters follow fairly well-known geologic and climate patterns. We did not train on 5 or 7 clusters. We aren't sure that is really within the scope of this paper, which has a fairly focused point about not training on a single basin.

l.249: "Fig ??" -> compilation error, please check reference

Response: Thank you.

Tables and Figures

Fig. 3.: y-axis label on right hand plot not necessary if figure outline kept like this

Response: Thank you.

Fig. 5 could be dropped since its content is also shown in Figure 6 and I think the value of showing the blue line alone is not as significant.

Response: That's true. We were really focused on readability, and thought the two plots helped isolate the main message from the "extra" experiments. But we see the reviewer's point and will take the suggestion.

Language

Very good and clear language, only small remarks as follow:

1. *63: "bounded bounded" -> bounded*
2. *69-70: "… in total 10 timesteps of streamflow observations…" -> … in total 10 streamflow observations…*

Response: Thank you.

Dear editor, dear SM,

We are very happy to see that the community comment option is used so extensively with respect to our manuscript. We decided to answer the comments of SM in a summarizing fashion, since many comments either overlap or are too off-topic.

Regarding expertise and writing style: We mainly would like to address these points to avoid the wrong impression appearing to future readers. We are experts on the use of LSTMs in hydrology. Not only did members of the group introduce the LSTM framework to rainfall–runoff modeling, but our group is also uniquely positioned in this regard because our members have done research both in traditional hydrological modeling and in Machine Learning. Two of the authors on this manuscript hold PhDs in hydrology and one author was formerly a professor of hydrology at multiple universities in the United States. The authors have over 50 peer-reviewed publications on the topic of hydrological modeling published in major hydrology journals, including several papers about the history and philosophy of hydrological modeling.

Further, given the other reviewers' overall feedback, we are quite happy with the tone, writing and style of our manuscript. We argue that the word "basin" is well-understood by the community and that the title and structure are geared towards providing the primary message of the paper. We believe that the active discussion and the comments of the other reviewers reflect the clarity of our writing. We will therefore not adapt any of the demanded changes in the community comments of SM with respect to writing style. Similarly, all the citations that we provide are correct, and we will adapt none of the related changes proposed in the community comments of SM.

Regarding modeling setup: In the comments there seems to be a confusion about the modeling setup in that some of the comments presume that we have a (semi) distributed setting, where multiple river reaches or HRU-like entities are routed to a single outlet. This is not the case. All models under investigation have a lumped setup. For example, when we say we use 200 basins to compare the performance of a single LSTM with 200 individual models, then each of these 200 basins is conceptually a headwater catchment with an associated gauging station. We are not sure what brought this confusion, but we will make sure that the revised manuscript mentions this explicitly.

Regarding the scope of the literature research: We agree with the overall assessment in the comments. The reason that we restricted the literature review to 2021 was that at the time of writing that seemed to be adequate and we wanted to have a manageable number of papers, however as time has progressed during the writing process, the submission, and the current review, we also believe that this literature review needs to be updated. We plan to expand the literature survey significantly in the revision.

—----------

Answers to Sivarajah Mylevaganam, 05 Feb 2024 CC1

I presume that the lifecycle of a manuscript goes through the steps outlined below.

1. *Performing a modeling work (e.g., mathematical modeling or physical modeling)*
2. *Writing a manuscript based what has been driven out from the modeling work.*
3. *Submitting the manuscript to a journal office that has advertised to have achieved a high impact factor.*
4. *Reviewing of the manuscript (probably a rigorous review) based on what has been presented in the manuscript by the authors.*
5. *Accepting or Rejecting based on what has been presented in the manuscript by the authors.*

Does step 4 assert that the authors have done the modeling work correctly? In my opinion, even a seasoned guy with extensive experience and many qualifications may not be able to catch a bug if the authors are smart enough to present the manuscripts in a way to get through the system of processing. The following questions would be pertinent to be asked at this level to understand the methodologies adopted by journal offices.

Is it possible to assert that there are no PhD students who have not tweaked modeling results or analysis to attain their titles and ambitions? Would we need hypothesis testing on this?
Is it possible to assert that faculty members who are on the edge of a knife have never tweaked modeling results or analysis to meet their needs?
Is it possible to assert that faculty members have never tweaked modeling results or analysis to progress in their careers?
Is it possible to assert that the faculty members who supervise students are conversant with all the nuts and bolts of the work carried out by the students?
Is it possible to assert that a manuscript accepted by a journal office with a high impact factor has no tweaked modeling results or analysis?

In my opinion, considering this loophole in the methodology adopted by journal offices, neither an advancement in the technologies nor in the scientific theories would yield the best of what is desirable for the betterment of the scientific field.

Response: All data, training code, trained models, and analysis notebooks were made public in the code and data repositories linked in the paper. Anyone can reproduce our results. Furthermore, all source datasets used in this paper are public and open-source and linked in the paper.

Line 90-Line 94:

We have some evidence that there might be ways to construct training sets that could result in better models than simply training on all available streamflow data. We do not have results that support this directly.

These statements are not understood.

Response: This statement comes from the attribute clustering results that are shown by the green line in Figure 6.

Line 79-Line 84: Figure 5 shows how test period performance increases as more basins are added to the training set. Performance continues to increase up to the maximum size of the CAMELS data set (531 basins). In other words, even these 531 basins are most likely not enough to train optimal LSTM models for streamflow.

In my opinion, these statements do not make sense. To better understand your methodology, assume that we have 16 basins (not 531 basins as in your analysis) as shown in the figure in the attached PDF file. For simplicity, let us forget the shapes of the basins. Moreover, assume that you have added the basins A, B, and C to have a training set to derive the associated NSE. Likewise you add more basins to your training set to have an array of NSE values to show a plot like the one that you have shown in Fig.5. Based on this figure, is it meaningful to conclude that the performance of the model increases as more basins are added to the training set? From a hydroloMylevaganamgical point of view, does it make sense to have a streamflow that is sourced by basins A, B, and C? Would the stream network that sources the flow at a location of interest become discontinuous?

Response: We are having trouble fully understanding this comment, however our reading of this comment (and based on other comments by this same reviewer) is that the reviewer is imagining a distributed hydrology model that operates over many subbasins connected by a channel network. However, the setup used in this paper is lumped catchment modeling. Based on the results in Figure 5 it is reasonable to conclude that deep learning streamflow models are generally better when trained on large datasets.

The NSE value that you have reported for a training set size of 100 basins (see Fig.5) may not actually represent the hydrology although the reported NSE value is very close to the NSE value for a training set size of 531 basins that may represent the actual hydrology. In ML models, you need to understand the theory that governs the system of equations. An in-depth understanding on the system of equations and how they are formulated will lead to understand the physics. Do the catchment attributes that you have chosen in your analysis play a role in the NSE values that you have reported in Fig.5?

Response: Catchment attributes strongly influence predictability – an example of that type of analysis can be found in this preprint: https://arxiv.org/abs/2307.16104. We are not interested in entertaining a discussion about the relative merits of physics vs. machine learning, as that topic has been covered extensively in prior literature and is largely irrelevant for the purpose of this paper.

11) Line 12-Line 13:

Because ML models are trained with data from multiple watersheds, they are able to learn hydrologically diverse rainfall–runoff responses (Kratzert et al., 2019b) in a way that is useful for example for prediction in ungauged basins (Kratzert et al., 2019a).

First In First Out (FIFO) is the concept that is implemented in citing (i.e., 2019a should come first). Since I don't know what is being implemented by this journal office, I would let the journal office to pay an attention on this.

Response: These two references are ordered alphabetically by title, according to APA guidelines. Please notice that this ordering style was implemented by the Copernicus journal group in the Latex template that they provide to authors, and was not chosen by the authors themselves.

12) Line 21-Line 23:

Then, in the bottom-up approach, after a model is developed, we might work on regionalization strategies to extrapolate parameters and parameterizations to larger areas (e.g., Samaniego et al., 2010; Beck et al., 2016).

Using e.g., in citing previous research works is considered an evading language in the scientific field. This gives an indication that the authors have failed to document a comprehensive review of the literature.

Response: The reviewer is incorrect that citing selected literature as examples of a broader set of papers is improper.

13) Line 34-Line 36:

We collected these 100 papers for review in September, 2022, nearly three years after the original regional LSTM rainfall–runoff modeling papers (Kratzert et al., 2019a, b) were published.

The critical review of these 100 papers is not found the current version of the manuscript. Throughout this manuscript, the authors have freely cited their own works. Considering the level of their knowledge in the field of hydrology and the other related disciplines that are reflected in the current version of the manuscript, I would have my reservation in the cited manuscripts.

Response: The majority of these 100 papers made a serious mistake, and this paper is focused on explaining what that mistake is and why it is important. We therefore do not and would not reference these papers as examples of background literature in the normal way.

13) Line 106-Line 107:

We selected a k-means cluster model based on a maximin criterion on silhouette scores, which resulted in a model with 6 clusters ranging from 59 to 195 basins per cluster.

I would say that your clustering methodology is inappropriate for this task as it would completely destroy the stream network and the underlying hydrology. This is one of the reasons why we have HUCs in the datasets that you have used in your analysis. Do you know the exact definition of HUC and the rationale behind developing HUCs? Your Fig.A4 is completely meaningless considering the purpose of the manuscript. I would suggest you have an in-depth look at your Fig.A3 and Fig.A4. What do you learn from those figures?

Response: Notice from Figure 5 that attributes-based clustering yielded better results than HUC-based grouping. We believe that what the reviewer is implying about Figures A3 and A4 is that attributes-based clusters show an element of geographical locality or organization. What we can infer from that is that climate, soils, and vegetation properties have non-random geographical distributions, which is well-known.

14) Line 235-Line 240:

For a particular basin of your interest, would you be able to show the values of Ws and the associated hs?

Response: The weights and biases of the model are the same for all basins. Again, this is the main point of this manuscript, which the reviewer seems to have missed. There are tens of thousands of weights and biases in the trained model, and it is not possible to show these values in text format in the paper. The trained models, including weights and biases, are included in the data repository referenced in the paper.

Answers to Sivarajah Mylevaganam, 06 Feb 2024 CC2

Sivarajah Mylevaganam

Alumnus, Spatial Sciences Laboratory, Texas A&M University, College Station, USA.

> The history of hydrological models goes to many decades. The progress that has been made to improve hydrological models through scientific findings has been showing the endless road to guide the next generation of hydrologist and associated specialists to quest for the betterment and mark the dead end. In this manuscript, the authors, who are not from the same hydrological basin, have employed an ML model to predict streamflow by training the model with hydrologically diverse data that is spatially and temporally large in extent. Based on the research, the authors draw a concrete conclusion that the previous modeling works (that is found based on ad-sense free search) using ML models have failed to understand the underlying philosophy in training and testing ML models.
>
> In my opinion, the current version of the manuscript has many flaws. Moreover, the way the manuscript has been presented gives an impression that the authors are far off from the field of hydrology. Therefore, the current version of the manuscript needs an expert in hydrology to go through in detail in an unbiased manner. Furthermore, the language that has been used in the manuscript needs to be edited by a language specialist as the way the manuscript has been presented and the words that have been coined throughout the manuscripts are beyond from what is expected in a scientific manuscript that is submitted to a journal office that advertises to have achieved a high impact factor based on rigorous reviews by a panel of experts in the fields of specializations listed in the scope of the journal.

Response: Two of the authors on this manuscript hold PhDs in hydrology and one author was formerly a professor of hydrology at multiple universities in the United States. The authors have over 50 peer-reviewed publications on the topic of hydrological modeling published in major hydrology journals, including several papers about the history and philosophy of hydrological modeling.

> In my opinion, the abstract of the manuscript needs to be re-written. I would say that an ML model with limited and poor data has been employed in writing the abstract. Moreover, what has been highlighted (“there is one MISTAKE in particular that is common”) is not well documented.

Response: The sentence that the reviewer cites answers this question directly. The full sentence reads: “*A large majority of studies that use this type of model do not follow best practices, and there is one mistake in particular that is common: training deep learning models on small, homogeneous data sets (i.e., data from one or a small number of*

watersheds)". We do not intend to rewrite the abstract substantively during the revision process.

> Line 10-Line 14:
>
> Hydrology models based on machine learning (ML) are different – ML models work best when trained on data from many watersheds (Nearing et al., 2021). This citation needs to be evaluated against the conclusion (i.e., ML models are best when trained with a large amount of hydrologically diverse data) that the authors draw from the manuscript that is submitted to this journal office.

Response: The Nearing et al. (2021) paper is in agreement with the conclusion of the current manuscript. We know this because we wrote both papers.

> Line 10-Line 14:
>
> Because ML models are trained with data from multiple watersheds, they are able to learn hydrologically diverse rainfall–runoff responses (Kratzert et al., 2019b). This citation goes to 2019. The comment 2 goes to 2021. Are the cited manuscripts giving the same thoughts? If so, the rationales in citing these manuscripts are not well understood.

Response: The 2021 paper cited in the first sentence gives the main results of this comparison. The 2019 papers cited in the second sentence give specific background context to this finding.

> Line 17-Line 23:
>
> The paragraph needs to be critically reviewed by a specialist. The sentences need to be evaluated. The paragraph gives an impression that the authors lack fundamental knowledge in the subject. The terminologies and words from an English dictionary are thrown without understanding the exact meaning.

Response: The authors are fluent in English and are experts in the subject of hydrological modeling. It would be helpful if the reviewer provided specific comments about what they don't understand or think that we misunderstand.

> The crux of the manuscript that is highlighted by the authors (i.e., LSTM stream models are best when trained with a large amount of hydrologically diverse data) is a well established fact in the scientific field. Basically, the authors are hitting the concept of SAMPLING SIZE of an experiment. Therefore, instead of going through a painful path of running models to determine the number of basins, it would be a wise man thought to go through some statistical methods to answer this question (i.e., sampling size). In fact, even the current version of this manuscript does not give the exact number that would be required. It is a random number (531) that the authors have ended up with based on what has been analyzed (see Fig.5).
> Line 30-Line 36:

Response: This paper is not about the sample size of an experiment, it is about the size of a training dataset for a machine learning model – these are not the same thing. The reviewer's suggestion would not address the central question that this paper deals with.

The choice of 531 basins is not random. This is a subset of the CAMELS US dataset (Newman et al. 2015, Addor et al. 2017) that was proposed by the authors of that dataset as a valid subset of basins for benchmarking hydrology models. This dataset is commonly used for community benchmarking of hydrology models.

> Does the order of the KEYWORDS have an influence in your search? Is the search from the engine not prioritized by the engine provider based on the business model employed? What was the reason to limit the search to 2021. Based on Fig.1, it is understood that there are more than 3500 publications. Even if we consider the authors' statement that review was initiated in September 2022, an iota of incompleteness surface.

Response: We plan to expand this literature survey significantly in the revision. No, the order of the keywords doesn't change the results on Google Scholar, at least not according to our observation. The reason that we restricted the literature review to 2021 was to have a manageable number of papers, however as time has progressed since writing, submitting, and now reviewing this manuscript, this literature review needs to be updated.

> Line 25-Line 29:
>
> The authors claim that the use of LSTMs for rainfall–runoff modeling has increased exponentially in the last several years. A figure (Fig.1) to support the claim is found in the manuscript. As per the figure, a rough estimation considering the heights of the bars gives an indication that around 8500(=3500+2500+1500+500+400+100) manuscripts have been found in the topic that the authors have invested. Referring to the previous comment, the authors have considered 100 manuscripts based on the search from a search engine of their interest. In other words, this manuscript is based on 100/8500*100%=1% of the manuscript found in the literature. What can be inferred from the training dataset that is employed in reviewing the literature? Will it lead to conclude that similar to an ML model the limited publications reviewed lead to draw wrong conclusions?

Response: As stated above, the literature review will be extended for the revision.

> The title of the manuscript needs to be assessed by a specialist. What is a basin? What is a watershed? What is a catchment? What is a region? What is the amount of data that a single basin possesses? What is the spatial and temporal extent of the basin that the authors are defining in the title of the manuscript? What is the heterogeneity level of the basin that the authors are defining in the title of the manuscript?
> Refer to Part III

Response: We believe that the word “basin” is well-understood by most hydrologists and that the title concisely summarizes the primary message of the paper.

Answers to Sivarajah Mylevaganam, 06 Feb 2024 CC3

Alumnus, Spatial Sciences Laboratory, Texas A&M University, College Station, USA.

9) Line 90-Line 94:

> We have some evidence that there might be ways to construct training sets that could result in better models than simply training on all available streamflow data. We do not have results that support this directly.
>
> These statements are not understood.

Response: This is a fairly simple and straightforward message. We are unsure how to make these sentences more clear.

> 10) Line 79-Line 84: Figure 5 shows how test period performance increases as more basins are added to the training set. Performance continues to increase up to the maximum size of the CAMELS data set (531 basins). In other words, even these 531 basins are most likely not enough to train optimal LSTM models for streamflow.
>
> In my opinion, these statements do not make sense. To better understand your methodology, assume that we have 16 basins (not 531 basins as in your analysis) as shown in the figure in the attached PDF file. For simplicity, let us forget the shapes of the basins. Moreover, assume that you have added the basins A, B, and C to have a training set to derive the associated NSE. Likewise you add more basins to your training set to have an array of NSE values to show a plot like the one that you have shown in Fig.5. Based on this figure, is it meaningful to conclude that the performance of the model increases as more basins are added to the training set? From a hydrological point of view, does it make sense to have a streamflow that is sourced by basins A, B, and C? Would the stream network that sources the flow at a location of interest become discontinuous?
>
> The NSE value that you have reported for a training set size of 100 basins (see Fig.5) may not actually represent the hydrology although the reported NSE value is very close

to the NSE value for a training set size of 531 basins that may represent the actual hydrology. In ML models, you need to understand the theory that governs the system of equations. An in-depth understanding on the system of equations and how they are formulated will lead to understand the physics. Do the catchment attributes that you have chosen in your analysis play a role in the NSE values that you have reported in Fig.5?

Response: We do not agree with the reviewer's assessment of the results in Figure 5. As mentioned previously, we are not interested in discussing the relative merits of machine learning vs. process-based hydrological modeling either in the paper or in these review comments, and we refer the reviewer to our previous publications if they are interested in our thoughts about this difference.

Acknowledgement and Disclaimer

The author is an alumnus of Texas A&M University, Texas, USA. The views expressed here are solely those of the author in his private capacity and do not in any way represent the views of Texas A&M University, Texas, USA.

Answers to Sivarajah Mylevaganam, 07 Feb 2024 CC4

Alumnus, Spatial Sciences Laboratory, Texas A&M University, College Station, USA.

11) Line 12-Line 13:

Because ML models are trained with data from multiple watersheds, they are able to learn hydrologically diverse rainfall–runoff responses (Kratzert et al., 2019b) in a way that is useful for example for prediction in ungauged basins (Kratzert et al., 2019a).

First In First Out (FIFO) is the concept that is implemented in citing (i.e., 2019a should come first). Since I don't know what is being implemented by this journal office, I would let the journal office to pay an attention on this.

Response: This comment is reproduced from one of this reviewer's previous submissions. The reviewer is incorrect about the ordering of references with the same author and year.

12) Line 21-Line 23:

Then, in the bottom-up approach, after a model is developed, we might work on regionalization strategies to extrapolate parameters and parameterizations to larger areas (e.g., Samaniego et al., 2010; Beck et al., 2016).

Using e.g., in citing previous research works is considered an evading language in the scientific field. This gives an indication that the authors have failed to document a comprehensive review of the literature.

Response: This comment is reproduced from one of this reviewer's previous submissions. The reviewer is incorrect about citing examples being evading or inappropriate.

13) Line 34-Line 36:

We collected these 100 papers for review in September, 2022, nearly three years after the original regional LSTM rainfall–runoff modeling papers (Kratzert et al., 2019a, b) were published.

The critical review of these 100 papers is not found the current version of the manuscript. Throughout this manuscript, the authors have freely cited their own works. Considering the level of their knowledge in the field of hydrology and the other related disciplines that are reflected in the current version of the manuscript, I would have my reservation in the cited manuscripts.

Response: This comment is reproduced from one of this reviewer's previous submissions. Please see our response above.

13) Line 106-Line 107:

We selected a k-means cluster model based on a maximin criterion on silhouette scores, which resulted in a model with 6 clusters ranging from 59 to 195 basins per cluster.

I would say that your clustering methodology is inappropriate for this task as it would completely destroy the stream network and the underlying hydrology. This is one of the reasons why we have HUCs in the datasets that you have used in your analysis. Do you know the exact definition of HUC and the rationale behind developing HUCs? Your Fig.A4 is completely meaningless considering the purpose of the manuscript. I would suggest you have an in-depth look at your Fig.A3 and Fig.A4. What do you learn from those figures?

Response: We used a lumped basin approach. There is no routing model or streamflow network in our modeling approach.

14) Line 235-Line 240:

For a particular basin of your interest, would you be able to show the values of Ws and the associated hs?

Response:

Acknowledgement and Disclaimer

The author is an alumnus of Texas A&M University, Texas, USA. The views expressed here are solely those of the author in his private capacity and do not in any way represent the views of Texas A&M University, Texas, USA.

Answers to Sivarajah Mylevaganam, 09 Feb 2024 CC5

Does the number of basins influence the modeling outcome using ML models?

A systematic methodology to answer the question that is being raised is presented. This method ensures that the underlying hydrology is conserved at the best to answer the question.

Step-1:

Draw an artificial stream network. The stream network shall have one limb or line per basin. In other words, if you have 531 basins in your dataset, your stream network will have 531 limbs or lines.

Step-2:

Use one of the existing techniques to order or name the stream developed in step 1. For example, if you are using Strahler's method to order your stream network, you will end up with a network like the one shown below. There are many methods to order or name a stream network. Therefore, you will have to do some research to determine the best method that is suited for the problem that you are solving.

Step-3:

Run your ML model, considering all the basins that you have in your dataset. Report the coefficient that is of your interest (e.g., NSE).

Step-4:

Run your ML model, considering the basins after removing the smallest stream number. If you are referring to the above figure, the smallest number will be 1. You can randomly remove the streams from the network.

Step-4-1:

Assume that you have 50 basins with "1". You can run your ML model after removing those 50 basins. In other words, you will have 531–50=481 basins in your simulation. Report the coefficient that is of your interest (e.g., NSE).

Step-4-2:

Assume that you have 50 basins with “1”. You can run your ML model after removing one of those 50 basins. You can randomly remove the basin from the network. In other words, you will have 531-1=530 basins in your simulation. You repeat this procedure by removing them one by one. In other words, your final simulation will have 531–50=481 basins in your simulation. Report the coefficient that is of your interest (e.g., NSE). You will have an array of NSEs.

Step-5:

Run your ML model, considering the basins after removing the next smallest stream number. If you are referring to the above figure, the next smallest number will be 2. You can randomly remove the streams from the network.

Step-5-1: Refer to Step-4-1

Step-5-2: Refer to Step-4-2

Step-6:

Continue your simulation work until you reach the largest stream number in your network. If you are referring to the above figure, the largest stream number will be 3.

Step-7:

Plot your NSEs and answer the question that has been raised.

In summary, a reverse algorithm needs to be developed to address the problem. Moreover, being conversant with spatial operations using products from ESRI or other vendors is required.

Answers to Sivarajah Mylevaganam, 06 Feb 2024 CC3

Alumnus, Spatial Sciences Laboratory, Texas A&M University, College Station, USA.

9) Line 90-Line 94:

> We have some evidence that there might be ways to construct training sets that could result in better models than simply training on all available streamflow data. We do not have results that support this directly.
>
> These statements are not understood.

Response: This is a fairly simple and straightforward message. We are unsure how to make these sentences more clear.

> 10) Line 79-Line 84: Figure 5 shows how test period performance increases as more basins are added to the training set. Performance continues to increase up to the maximum size of the CAMELS data set (531 basins). In other words, even these 531 basins are most likely not enough to train optimal LSTM models for streamflow.
>
> In my opinion, these statements do not make sense. To better understand your methodology, assume that we have 16 basins (not 531 basins as in your analysis) as shown in the figure in the attached PDF file. For simplicity, let us forget the shapes of the basins. Moreover, assume that you have added the basins A, B, and C to have a training set to derive the associated NSE. Likewise you add more basins to your training set to have an array of NSE values to show a plot like the one that you have shown in Fig.5. Based on this figure, is it meaningful to conclude that the performance of the model increases as more basins are added to the training set? From a hydrological point of view, does it make sense to have a streamflow that is sourced by basins A, B, and C? Would the stream network that sources the flow at a location of interest become discontinuous?
>
> The NSE value that you have reported for a training set size of 100 basins (see Fig.5) may not actually represent the hydrology although the reported NSE value is very close to the NSE value for a training set size of 531 basins that may represent the actual hydrology. In ML models, you need to understand the theory that governs the system of equations. An in-depth understanding on the system of equations and how they are formulated will lead to understand the physics. Do the catchment attributes that you have chosen in your analysis play a role in the NSE values that you have reported in Fig.5?

Response: We do not agree with the reviewer's assessment of the results in Figure 5, and we believe that the reviewer may have fundamentally misunderstood the experiments shown in this figure (as explained in our previous responses to this reviewer).

Furthermore, as mentioned previously, we are not interested in discussing the relative merits of machine learning vs. process-based hydrological modeling either in the paper or in these review comments, and we refer the reviewer to our previous publications if they are interested in our thoughts about that distinction.

Acknowledgement and Disclaimer

The author is an alumnus of Texas A&M University, Texas, USA. The views expressed here are solely those of the author in his private capacity and do not in any way represent the views of Texas A&M University, Texas, USA.

Answers to Sivarajah Mylevaganam, 07 Feb 2024 CC4

Alumnus, Spatial Sciences Laboratory, Texas A&M University, College Station, USA.

11) Line 12-Line 13:

Because ML models are trained with data from multiple watersheds, they are able to learn hydrologically diverse rainfall–runoff responses (Kratzert et al., 2019b) in a way that is useful for example for prediction in ungauged basins (Kratzert et al., 2019a).

First In First Out (FIFO) is the concept that is implemented in citing (i.e., 2019a should come first). Since I don't know what is being implemented by this journal office, I would let the journal office to pay an attention on this.

Response: This comment is reproduced from one of this reviewer's previous submissions. The reviewer is incorrect about the ordering of references with the same author and year.

12) Line 21-Line 23:

Then, in the bottom-up approach, after a model is developed, we might work on regionalization strategies to extrapolate parameters and parameterizations to larger areas (e.g., Samaniego et al., 2010; Beck et al., 2016).

Using e.g., in citing previous research works is considered an evading language in the scientific field. This gives an indication that the authors have failed to document a comprehensive review of the literature.

Response: This comment is reproduced from one of this reviewer's previous submissions. The reviewer is incorrect about citing examples being evading or inappropriate.

13) Line 34-Line 36:

We collected these 100 papers for review in September, 2022, nearly three years after the original regional LSTM rainfall–runoff modeling papers (Kratzert et al., 2019a, b) were published.

The critical review of these 100 papers is not found the current version of the manuscript. Throughout this manuscript, the authors have freely cited their own works. Considering the level of their knowledge in the field of hydrology and the other related disciplines that are reflected in the current version of the manuscript, I would have my reservation in the cited manuscripts.

Response: This comment is reproduced from one of this reviewer's previous submissions. Please see our response above.

13) Line 106-Line 107:

We selected a k-means cluster model based on a maximin criterion on silhouette scores, which resulted in a model with 6 clusters ranging from 59 to 195 basins per cluster.

I would say that your clustering methodology is inappropriate for this task as it would completely destroy the stream network and the underlying hydrology. This is one of the reasons why we have HUCs in the datasets that you have used in your analysis. Do you know the exact definition of HUC and the rationale behind developing HUCs? Your Fig.A4 is completely meaningless considering the purpose of the manuscript. I would suggest you have an in-depth look at your Fig.A3 and Fig.A4. What do you learn from those figures?

Response: We used a lumped basin approach. There is no routing model or streamflow network in our modeling approach.

14) Line 235-Line 240:

For a particular basin of your interest, would you be able to show the values of Ws and the associated hs?

Response:

Acknowledgement and Disclaimer

The author is an alumnus of Texas A&M University, Texas, USA. The views expressed here are solely those of the author in his private capacity and do not in any way represent the views of Texas A&M University, Texas, USA.

Answers to Sivarajah Mylevaganam, 09 Feb 2024 CC5

Does the number of basins influence the modeling outcome using ML models?

A systematic methodology to answer the question that is being raised is presented. This method ensures that the underlying hydrology is conserved at the best to answer the question.

Step-1:

Draw an artificial stream network. The stream network shall have one limb or line per basin. In other words, if you have 531 basins in your dataset, your stream network will have 531 limbs or lines.

Step-2:

Use one of the existing techniques to order or name the stream developed in step 1. For example, if you are using Strahler's method to order your stream network, you will end up with a network like the one shown below. There are many methods to order or name a stream network. Therefore, you will have to do some research to determine the best method that is suited for the problem that you are solving.

Step-3:

Run your ML model, considering all the basins that you have in your dataset. Report the coefficient that is of your interest (e.g., NSE).

Step-4:

Run your ML model, considering the basins after removing the smallest stream number. If you are referring to the above figure, the smallest number will be 1. You can randomly remove the streams from the network.

Step-4-1:

Assume that you have 50 basins with "1". You can run your ML model after removing those 50 basins. In other words, you will have 531–50=481 basins in your simulation. Report the coefficient that is of your interest (e.g., NSE).

Step-4-2:

Assume that you have 50 basins with “1”. You can run your ML model after removing one of those 50 basins. You can randomly remove the basin from the network. In other words, you will have 531-1=530 basins in your simulation. You repeat this procedure by removing them one by one. In other words, your final simulation will have 531–50=481 basins in your simulation. Report the coefficient that is of your interest (e.g., NSE). You will have an array of NSEs.

Step-5:

Run your ML model, considering the basins after removing the next smallest stream number. If you are referring to the above figure, the next smallest number will be 2. You can randomly remove the streams from the network.

Step-5-1: Refer to Step-4-1

Step-5-2: Refer to Step-4-2

Step-6:

Continue your simulation work until you reach the largest stream number in your network. If you are referring to the above figure, the largest stream number will be 3.

Step-7:

Plot your NSEs and answer the question that has been raised.

In summary, a reverse algorithm needs to be developed to address the problem. Moreover, being conversant with spatial operations using products from ESRI or other vendors is required.

Answers to Sivarajah Mylevaganam, 07 Feb 2024 CC4

Alumnus, Spatial Sciences Laboratory, Texas A&M University, College Station, USA.

11) Line 12-Line 13:

Because ML models are trained with data from multiple watersheds, they are able to learn hydrologically diverse rainfall–runoff responses (Kratzert et al., 2019b) in a way that is useful for example for prediction in ungauged basins (Kratzert et al., 2019a).

First In First Out (FIFO) is the concept that is implemented in citing (i.e., 2019a should come first). Since I don't know what is being implemented by this journal office, I would let the journal office to pay an attention on this.

Response: This comment is reproduced from one of this reviewer's previous submissions. References are ordered according to APA style guidelines, and also according to the Latex template supplied by the journal.

12) Line 21-Line 23:

Then, in the bottom-up approach, after a model is developed, we might work on regionalization strategies to extrapolate parameters and parameterizations to larger areas (e.g., Samaniego et al., 2010; Beck et al., 2016).

Using e.g., in citing previous research works is considered an evading language in the scientific field. This gives an indication that the authors have failed to document a comprehensive review of the literature.

Response: This comment is reproduced from one of this reviewer's previous submissions. The reviewer is incorrect about citing examples being evading or inappropriate.

13) Line 34-Line 36:

We collected these 100 papers for review in September, 2022, nearly three years after the original regional LSTM rainfall–runoff modeling papers (Kratzert et al., 2019a, b) were published.

The critical review of these 100 papers is not found the current version of the manuscript. Throughout this manuscript, the authors have freely cited their own works. Considering the level of their knowledge in the field of hydrology and the other related disciplines that are reflected in the current version of the manuscript, I would have my reservation in the cited manuscripts.

Response: This comment is reproduced from one of this reviewer's previous submissions. Please see our response to the original.

13) Line 106-Line 107:

We selected a k-means cluster model based on a maximin criterion on silhouette scores, which resulted in a model with 6 clusters ranging from 59 to 195 basins per cluster.

I would say that your clustering methodology is inappropriate for this task as it would completely destroy the stream network and the underlying hydrology. This is one of the reasons why we have HUCs in the datasets that you have used in your analysis. Do you know the exact definition of HUC and the rationale behind developing HUCs? Your Fig.A4 is completely meaningless considering the purpose of the manuscript. I would suggest you have an in-depth look at your Fig.A3 and Fig.A4. What do you learn from those figures?

Response: We used a lumped basin approach. There is no routing model or streamflow network in our modeling approach.

14) Line 235-Line 240:

For a particular basin of your interest, would you be able to show the values of Ws and the associated hs?

Response: The reviewer asked this question in different wording in a previous comment. Please see our original response.

Acknowledgement and Disclaimer

The author is an alumnus of Texas A&M University, Texas, USA. The views expressed here are solely those of the author in his private capacity and do not in any way represent the views of Texas A&M University, Texas, USA.

Answers to Sivarajah Mylevaganam, 09 Feb 2024 CC5

Does the number of basins influence the modeling outcome using ML models?

A systematic methodology to answer the question that is being raised is presented. This method ensures that the underlying hydrology is conserved at the best to answer the question.

Step-1:

Draw an artificial stream network. The stream network shall have one limb or line per basin. In other words, if you have 531 basins in your dataset, your stream network will have 531 limbs or lines.

Step-2:

Use one of the existing techniques to order or name the stream developed in step 1. For example, if you are using Strahler's method to order your stream network, you will end up with a network like the one shown below. There are many methods to order or name a stream network. Therefore, you will have to do some research to determine the best method that is suited for the problem that you are solving.

Step-3:

Run your ML model, considering all the basins that you have in your dataset. Report the coefficient that is of your interest (e.g., NSE).

Step-4:

Run your ML model, considering the basins after removing the smallest stream number. If you are referring to the above figure, the smallest number will be 1. You can randomly remove the streams from the network.

Step-4-1:

Assume that you have 50 basins with "1". You can run your ML model after removing those 50 basins. In other words, you will have 531–50=481 basins in your simulation. Report the coefficient that is of your interest (e.g., NSE).

Step-4-2:

Assume that you have 50 basins with “1”. You can run your ML model after removing one of those 50 basins. You can randomly remove the basin from the network. In other words, you will have 531-1=530 basins in your simulation. You repeat this procedure by removing them one by one. In other words, your final simulation will have 531–50=481 basins in your simulation. Report the coefficient that is of your interest (e.g., NSE). You will have an array of NSEs.

Step-5:

Run your ML model, considering the basins after removing the next smallest stream number. If you are referring to the above figure, the next smallest number will be 2. You can randomly remove the streams from the network.

Step-5-1: Refer to Step-4-1

Step-5-2: Refer to Step-4-2

Step-6:

Continue your simulation work until you reach the largest stream number in your network. If you are referring to the above figure, the largest stream number will be 3.

Step-7:

Plot your NSEs and answer the question that has been raised.

In summary, a reverse algorithm needs to be developed to address the problem. Moreover, being conversant with spatial operations using products from ESRI or other vendors is required.

Answers to Sivarajah Mylevaganam, 09 Feb 2024 CC5

Does the number of basins influence the modeling outcome using ML models?

A systematic methodology to answer the question that is being raised is presented. This method ensures that the underlying hydrology is conserved at the best to answer the question.

Step-1: Draw an artificial stream network. The stream network shall have one limb or line per basin. In other words, if you have 531 basins in your dataset, your stream network will have 531 limbs or lines.

Step-2: Use one of the existing techniques to order or name the stream developed in step 1. For example, if you are using Strahler's method to order your stream network, you will end up with a network like the one shown below. There are many methods to order or name a stream network. Therefore, you will have to do some research to determine the best method that is suited for the problem that you are solving.

Step-3: Run your ML model, considering all the basins that you have in your dataset. Report the coefficient that is of your interest (e.g., NSE).

Step-4: Run your ML model, considering the basins after removing the smallest stream number. If you are referring to the above figure, the smallest number will be 1. You can randomly remove the streams from the network.

Step-4-1: Assume that you have 50 basins with "1". You can run your ML model after removing those 50 basins. In other words, you will have 531–50=481 basins in your simulation. Report the coefficient that is of your interest (e.g., NSE).

Step-4-2: Assume that you have 50 basins with “1”. You can run your ML model after removing one of those 50 basins. You can randomly remove the basin from the network. In other words, you will have 531-1=530 basins in your simulation. You repeat this procedure by removing them one by one. In other words, your final simulation will have 531–50=481 basins in your simulation. Report the coefficient that is of your interest (e.g., NSE). You will have an array of NSEs.

Step-5: Run your ML model, considering the basins after removing the next smallest stream number. If you are referring to the above figure, the next smallest number will be 2. You can randomly remove the streams from the network.

Step-5-1: Refer to Step-4-1

Step-5-2: Refer to Step-4-2

Step-6: Continue your simulation work until you reach the largest stream number in your network. If you are referring to the above figure, the largest stream number will be 3.

Step-7: Plot your NSEs and answer the question that has been raised.

In summary, a reverse algorithm needs to be developed to address the problem. Moreover, being conversant with spatial operations using products from ESRI or other vendors is required.

Response: We do not use a stream network in our experiments. Our experiments use a lumped catchment model. These experiments sound interesting, but not relevant to the purpose of this paper.

Answers to Sivarajah Mylevaganam, 28 Feb 2024 CC5

Does the number of catchments/watersheds/sub-watersheds/subbasins influence the modeling outcome using ML models?

A systematic methodology to answer the question that is being raised is presented. This method ensures that the underlying hydrology is conserved at the best to answer the question.

Step-1:

Identify the basin of your interest. This should be based on a proper research methodology (Refer to PART IX)

Step-2:

Extract the associated catchments and the stream network using ESRI products or other vendors. These catchments and the stream network cover the basin of your interest. See the attached PDF file.

Step-3:

Use one of the existing techniques to order or name the stream network developed in step 2. For example, if you are using Strahler's method to order your stream network, you will end up with a network like the one shown below. There are many methods to order or name a stream network. Therefore, you will have to do some research to determine the best method that is suited for the problem that you are solving.

Step-4:

Run your ML model, considering all the catchments that you have in your basin. Report the coefficient that is of your interest (e.g., NSE).

Step-5:

Run your ML model, considering the catchments after removing the smallest stream number. If you are referring to the above figure, the smallest number will be 1. You can randomly remove the streams from the network.

Step-5-1:

Assume that you have 500 catchments with "1". Assume that you have 10,000 catchments within your basin. You can run your ML model after removing those 500 catchments. In other words, you will have 10,000–500=9,500 catchments in your simulation. Report the coefficient that is of your interest (e.g., NSE).

Step-5-2:

Assume that you have 500 catchments with "1". Assume that you have 10,000 catchments within your basin. You can run your ML model after removing one of those 500 catchments. You can randomly remove the catchment from the network. In other words, you will have 10,000-1=9,999 catchments in your simulation. You repeat this procedure by removing them one by one. In other words, your final simulation will have 10,000–500=9,500 catchments in your simulation. Report the coefficient that is of your interest (e.g., NSE). You will have an array of NSEs.

Step-6:

Run your ML model, considering the catchments after removing the next smallest stream number. If you are referring to the above figure, the next smallest number will be 2. You can randomly remove the streams from the network.

Step-6-1: Refer to Step-5-1

Step-6-2: Refer to Step-5-2

Step-7:

Continue your simulation work until you reach the largest stream number in your network. If you are referring to the above figure, the largest stream number will be 3.

Step-8:

Plot your NSEs and answer the question that has been raised.

In summary, a reverse algorithm needs to be developed to address the problem. Moreover, being conversant with spatial operations using products from ESRI or other vendors is required.

Response: This is not an appropriate methodology to address the question that is posed in the paper under review. The reviewer fundamentally misunderstands what was done in this paper. This is now the 6th set of comments submitted by this reviewer, and given the large disconnect between this reviewer's comments and the actual subject matter of the paper, we are unable to respond in more substantive detail.

Comments by M. Harchowitz in *blue*, our responses in black.

In the manuscript "Never train an LSTM on a single basin" by Krazert et al., the authors convincingly address the very important subject of confronting hydrological models for training/calibration with data that include a sufficiently large variability of environmental conditions. While it is widely acknowledged in the community that the "richness" of a training dataset plays a crucial role for identifying meaningful and robust models (read as: model architectures and parameters), the advent of powerful ML techniques, such as LSTMs has further exacerbated this issue and, as demonstrated by the authors, many studies do not fully (or not sufficiently well) exploit available data.

This manuscript is therefore a very welcome and probably even necessary reminder for the community to avoid being lured into questionable generalizations that may follow from insufficiently trained/tested ML models that are not actually supported by data. Overall, I find that this manuscript is built on an excellent level of reflection and is very well argued. I also believe that the clear focus on LSTMs, as emergent and powerful tool, is important and justified.

Response: Thank you!

Having said that, I nevertheless believe the manuscript could benefit from a somewhat wider view beyond LSTMs. Thus, while the focus on LSTMs is fine and important, I also believe that it would be helpful for the reader to project the LSTM focus onto a wider background canvas that, as a starting point, provides a more general modelling perspective as well as, here and there, more precise formulations with respect to process-based models (hereatier PB; including the entire spectrum from lumped conceptual to spatially explicit "physics-based" models).

Response: We disagree with the sentiment that including a wider discussion that includes PB models helps to improve a manuscript with a very clear focus on one thing: The correct way to train LSTMs in the context of hydrology. In our opinion, adding a discussion about general hydrological modeling in this manuscript would smear the focus that we want to discuss. There are dozens of papers in hydrology journals that cover the wider perspective on hydrological modeling – some of which have been written by the current authors themselves, some of which have been written by the reviewer himself.

As a baseline, ML approaches typically offer sufficient freedom and flexibility to identify the most efficient connection structure in a system, as next to the data fed into ML, in most cases (except for mass conserving ML models) little to no further mechanistic assumptions are imposed onto these models. This is their strength. In the theoretical case of "complete" knowledge, i.e. sufficient data, ML would without doubt be able to converge towards unique (and possibly time-variable) connection structures for each system (or catchment). The limiting factor here is, quite obviously, our lack of "complete" knowledge.

PB approaches, in contrast, typically impose very strict constraints on the functional architecture of models and thus on the connections in the system. This is done by imposing specific

parametric relationships that are meant to describe various storage/gradient – resistance processes in the system, which are known or assumed to be relevant in that specific system, but potentially not elsewhere. HOWEVER, in reality, these relationships are rarely or never known. In addition, these parametric relationships (e.g. linear reservoir as example for a very simple storage discharge relationship) are in most cases smooth and regular, while real world processes at scales larger than the lab-scale – mostly due to spatio-temporal heterogeneities in environmental conditions – need to be expected to be more jagged and irregular. From a historical perspective, PB models have indeed for a long time been developed and calibrated for specific locations. Applying these models to other catchments, using the same functional relationships (or at least relationships from the same family of functions/distributions) then frequently fails to reproduce the hydrological response elsewhere. That motivated the first attempts of flexibilization and customization of using modular PB model frameworks starting from Leavesley et al. (1996) up to more recent initiatives (e.g. Fenicia et al., 2011; Clark et al., 2015 and others). Other studies have demonstrated that allowing PB models more flexibility, either in terms process resolution and thus in the number of parameterized processes, spatial resolution or prior parameter distributions (e.g. Hrachowitz et al., 2014; Mendoza et al. 2015) can dramatically increase their performance, if at the same time balanced with more data to confront the model with. Related to that are of course also the many model regionalization attempts. The most successful so far is arguably the MPR scheme used in the mhM model (e.g. Samaniego et al., 2010), which the authors have cited in their manuscript. The culmination of the development so far is the recent paper by Gharari et al. (2021), in which it is argued that, in the absence of more detailed knowledge, the constraints of functional parametric relationships in PB models should in principle be relaxed to the point that they only have to satisfy mass conservation and the condition of being monotonic, which are essentially fundamental physical constraints. The training process of such a PB model would then, reflecting that of a ML approach, allow the model to flexibly generate and test parametric or potentially even non-parametric relationships, e.g. storage-discharge relationships, that are most consistent with available data.

Why did I now throw an almost one-page comparison of PB approaches at the authors? Because what the authors describe in their manuscript fundamentally applies to both, ML and PB, and should also be reflected at least in the context given in the introduction. From my perspective the only difference between ML and PB is the level of constraints (and thus assumed or real knowledge) imposed on the models: very low for ML, very high for PB. From the historical perspective PB has not been flexible due to (1) insufficient data before the availability of large sample and/or remote sensing datasets and (2) lack of computational capacity (in particular, for spatially explicit models). However, although so far it has not systematically been done, there is nothing to suggest that it could not be done. I therefore believe, it would be very valuable for the community if this clearly came across in the introduction that the value of data "richness" used for training is a general issue in modelling and not limited to ML. In the end, I am convinced that ML and PB models are merely two sides of the same medal (i.e. the observed hydrological system) and that eventually they will in their functionality converge towards each other.

Response: We partially agree, but then also mostly disagree. We agree that ML models and PB models can be seen on a spectrum of progressively higher inductive biases. However, we disagree with the reviewer's opinion that opening the discussion in this manuscript to speculative properties of PB models helps to improve this manuscript. As the reviewer mentions below, PB models do not currently behave in the way that the reviewer suggests that they might do at some point (i.e., by benefitting from calibration on large-sample data). It would be nice if they did, and we might even imagine finding a way to make that happen, but currently they simply don't behave that way. The results in the left-hand panel of Figure 2 use some of the parameter regionalization strategies that the reviewer mentions. Please notice that we did not create the data in the left-hand panel of Figure 2 – these model runs were done by the developers of the parametrization schemes who were (presumably) motivated to make their models and regionalization strategies perform as well as possible. The focus of this manuscript is to highlight the importance of the correct modeling setup when working with LSTMs and we would like to avoid distracting from this focus with discussions around possible (future) properties of PB models.

Specific comments:

p.1, l.8: conceptual models are a category of process-based models. Probably less ambiguous if "continuum-based" or "physics-based" used instead of "process-based"

Response: As far as we know there is no general consensus among hydrologists. As an example, when defining process-based models, Todini (2011) says the following:

"As an alternative to conceptual models several authors aimed at improving the physical representation of the rainfall-runoff process."

Our experience and perspective is that the process-based hydrological modeling community has worked very hard to draw a distinction between what they do and conceptual models. The latter being defined as models that generally don't explicitly represent hydrological processes, but instead use reduced-complexity approximations (while still incorporating certain physical laws like mass conservation). In other words, the relationship between conceptual, physically-based, and process-based hydrology models is, to our understanding, the opposite of what the reviewer suggests – conceptual models are a subset of physically-based models, but not a subset of process-based models.

Todini, Ezio. "History and perspectives of hydrological catchment modelling." *Hydrology Research* 42.2-3 (2011): 73-85.

p.1, l.9: this statement is not sufficiently precise and actually incorrect: PB models do not specifically require long data records. What they require instead is (as any model, one would

plausibly assume) sufficient data support. The difference being that the lengths of the records could just as well be balanced with the variety of data and loss functions, e.g. short time series can be complemented by multiple other time series of other variables, such as soil moisture, snow cover, groundwater levels, storage changes, evaporation, etc. (e.g. Nijzink et al., 2018; Dembélé et al., 2020; Hulsman et al., 2021). In the contrary, the use of long time series bears the risk of averaging out temporal variability in the model parameters, caused e.g. by natural or directly human-induced changes in vegetation (e.g. Hrachowitz et al., 2021; Tempel et al., 2024)

> **Response:** We will remove the reference to "long" data records and state that process-based models require calibration to observation data in the locations where they are applied in order to provide reasonable results.

p.1, l.10-16: I disagree. This is not a unique characteristic of ML. Flexibilize PB models and train them to multiple catchments will eventually converge to the same effects. In my opinion the difference is rather in the level of imposed constraints (see above). Thus, the fact that it is not yet done with PB models, does not mean that it cannot be done. I think it would be very helpful for the reader to make this difference clear here.

> **Response:** We think at this point the statement is speculation (see also our general reply above). At least for the moment, we consider it to be a fact that PB models are better when trained locally. This does not mean that we disagree that PB models could potentially benefit from regional training, but we see this kind of speculation out-of-scope for the focus of this manuscript.

p.1, l.17: not sure if "intuition" is the best term to use here

> **Response:** We actually think that "intuition" is the correct word here. In our opinion, the reason that we see the majority of applications of LSTMs in hydrology to single basins is because single-basin modeling is the *intuitive* thing to do for someone who studied hydrology and used to work with conceptual/process-based models. This always yielded the best results (and generally runs faster and is easier to set up). The point of this manuscript is that working with LSTMs requires us to change our default modeling approach (read: our intuition of what is good and what is not).

p.1, l.20: please see above: conceptual models are process-based models. In addition, conceptual models can be implemented at any spatial resolution from lumped, over semi-distributed to fully distributed (frequently referred to as data-gridded models then). Please adjust the statement.

> **Response:** For discussion around conceptual models and process-based models, see our reply above. Otherwise, we see no contradiction between the comment and what we wrote in the referenced lines, which does not imply that conceptual models have any limit to their spatial resolution.

p.1, l.20: I am not sure that in environmental sciences we can actually "verify" anything, given the uncertainties (or incomplete knowledge) in every part of the system. Perhaps better to rephrase to "test" or "evaluate"

> **Response:** We see what the reviewer is trying to point at, but we consider that a more of a philosophical discussion than what we actually wanted to discuss. We will still change this sentence to avoid that future will be diverted as the reviewer did. In general, "model verification", "model testing", and "model evaluation" are three different, well defined terms. Model verification defines the process of ensuring that the model is correctly implemented (i.e., no mistakes in the equations). Model testing refers to the process of probing the model robustness against, e.g., noise in the input data and to find the boundaries where the assumptions in a model break down. Model evaluation refers to the process of checking a (calibrated/trained) model on unseen data. In this regard, "verify" is probably not the correct word in this sentence, but for another reason. Therefore, we will change it to "evaluate" since this is what we meant here.

p.2, l.37: idem for PB models – nothing speaks against them being trained to a large sample either.

> **Response:** Nothing speaks against training PB models on large samples. However, as shown in Figure 2 it is not *beneficial* for PB models to be trained on multiple basins. To give another example: The results Mai et al. (2022) show that all PB models in that study perform worse when trained on multiple basins. Therefore, as of now, if one is concerned about model performance, there exist reasons why (some) PB models should indeed only be trained on a small number of watersheds (or even a single specific one).

*p.3, l.43: this does not really come as a surprise. The more variable and *rich* a data set is used for training the more robust the model. But should this not be true for any type of model in any discipline?*

> **Response:** Ideally it should, yes. But is it true for PB models? As discussed multiple times above, it is not reflected in the current state of PB models and regionalization schemes. For ML models, we agree that this should not be a surprise, yet a large majority of hydrologists train their ML models on tiny datasets, hence this manuscript. The requirement for large data logically follows from the conceptualization of (modern) ML models and might not be a new insight for some part of the community (like, e.g., the reviewer and most ML researchers). However, we believe that there exists a large subset of readers who will benefit from the explicit statement and we therefore plan to keep the lines as they are.

p.3, 46ff: Figure 2 is a great comparison that I have already found very useful when it was originally published a few years ago (Kratzert et al., 2019). However, what I have never managed to get my head around is the following: the PB models tested have between ~ 15

(VIC) and >50 (mhM) calibration parameters, although the calibration strategy does not become entirely clear from that paper. On the other hand, and apologies if I understand something wrong here, LSTMs are defined by a handful of hyperparameters, that regulate the number of actual trainable model parameters (or "weights" or any other jargon term that is equivalent to "parameters" in PB models). In my understanding and without looking up the input size used in the experiment underlying the Kratzert et al. (2019, 2021) analysis then leads me, assuming for the moment a lower limit of the input size as 1, using the following expression for the number of trainable parameters n = 4(Input size + Hidden size + 1)*Hidden size, to a bare minimum of 320 (Hidden size = 8 as reported in Appendix A2 and A3, p.10ff) or 264192 (Hidden size = 256 in Appendix A1, p.9ff) trainable parameters in the LSTMs used here. It leaves me profoundly confused, how models with such an elevated number of trainable parameters can be in a fair way compared to models that have at least one order of magnitude fewer parameters. This would be like comparing the time of a sprinter to the time of a person shackled in chains to finish a 100m race. For example, and although I have not tested this, I do not see a compelling reason why increasing the number of parameters in a PB model for calibration in a single basin from ~15 to >320, would not improve the model, plausibly even to the level of a LSTM trained for that basin. The same can of course be said for multi-basin calibration. As expressed above, the fact that standard PB models do not do that is different from the notion that they cannot do it. But again, I may be victim to a fundamental misunderstanding here. In any case, I would be glad to hear the authors perspective on that.*

Response: In our perspective, there is no concept of "fairness" related to the number of parameters, and especially not when comparing a conceptual/PB model, into which we hard-code our understanding of the underlying processes, with a neural network, which, without training, doesn't know anything. If hydrologists could build models with more parameters that work better, then by all means, they should do that. However, there are several studies that investigate this topic and come to a different conclusion than what is hypothesized here by the reviewer. For example, Perrin et al. (2001) write in section 7.5. (Does the number of free parameters increase model performances?) that "*In calibration mode, models with a larger number of parameters generally benefit from this increase in degrees of freedom and yield a better fit of observed data. But this trend disappears at the verification stage where models with a limited number of parameters achieve results as good as those of more complex models*". Orth et al. (2015) come to a similar conclusion: "*We conclude that added complexity does not necessarily lead to improved performance of hydrological models, and that performance can vary greatly depending on the considered hydrological variable (e.g. runoff vs. soil moisture) or hydrological conditions (floods vs. droughts)*". More recently, Merz et al. (2022) performed a large-sample study in which they varied model structures and number of free parameters and concluded that "*The results of our study favor simple model structures for large-scale applications, in which the main hydrological processes of runoff generation and routing in the soil layer, groundwater, and river network are conceptualized individually by a single storage*". That being said, ways to increase the performance of PB models to the level of LSTMs would certainly be a highly valuable finding. To conclude this point, we think that a discussion about the number of parameters between

LSTMs and PB models is out of scope for this manuscript that focuses on the correct use of LSTMs by themselves.

C. Perrin, C. Michel, V. Andréassian, “Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments” (2001). https://doi.org/10.1016/S0022-1694(00)00393-0

R. Orth, M. Staudinger, S. I. Seneviratne, J. Seibert, M. Zappa “Does model performance improve with complexity? A case study with three hydrological models” (2015) https://doi.org/10.1016/j.jhydrol.2015.01.044

R. Merz, A. Miniussi, S. Basso, K. Petersen, and L. Tarasova, 2022: “More Complex is Not Necessarily Better in Large-Scale Hydrological Modeling: A Model Complexity Experiment across the Contiguous United States” https://doi.org/10.1175/BAMS-D-21-0284.1

p.3, Figure 2 and captions thereof (but also Figures 5 and 6): NSE of what? I suppose stream flow Q. But please make sure to explicitly state that.

Response: Thank you, we will do that.

p.6, l.86: I would argue that volume and variety are not uncorrelated and that in the end, variety counts. This also seems to be the take away from Figure 6, where once variety is discounted for (e.g. attribute and HUC splits), volume does not really change the results. This suggests that volume does not really come into play.

Response: We are not sure if we agree with the assessment of Figure 6 and would even argue that it shows the opposite of what the reviewer has interpreted. As discussed in the manuscript (e.g., line 90ff), there is some evidence that homogenizing the training data *might* provide some value over purely random data splits. What is clear, however, is that increasing the training set size seems to always help, at least up to the limit of the 531 catchment explored in this paper. We will try to make this point clearer in the manuscript by linking Figure 6 to the description in l.90ff.

p.7, l.90ff: I completely agree. This has been shown in a considerable body of literature that demonstrates the beneficial effects of multi-objective, multi-criteria and/or multi-variable calibration with PB models going back to at least Gupta et al. (1998), and many studies since then (e.g. Hrachowitz et al., 2014; Nijzink et al., 2018; Dembélé et al., 2020; Hulsman et al., 2021a,b and many others). Why should this be different for ML approaches? Indeed, I am convinced that also LSTMs will benefit from such a multi-objective, - criteria or -variable approach.

Response: We are not sure how this comment is connected with the argument in l.90 ff.

p.7, l.128: not only LSTMs. All inverse model approaches require sufficiently "rich" data that allow to balance their flexibility (read: number of training parameters) with sufficient constraints, as argued e.g. by Gupta et al. (2008 and in particular Figure 4 therein; 2012) but in the end also by Kirchner (2006) and many others.

> **Response:** Given the scope of our manuscript, we are not sure that it makes sense to draw a direct parallel between intuitions about calibrating hydrology models and intuitions about training ML models. We prefer not to add this type of discussion to the paper.

p.7, 132: not sure I fully understand this statement. Did not some recent papers that were partly coauthored by some of the authors provide the first steps in "adding physics" to LSTMs by enforcing conservation of mass and/or energy (e.g. Hoedt et al., 2021; Frame et al., 2023; Pokharel et al., 2023)?

> **Response:** It is correct that we also co-authored papers that tried to add physics into machine learning models (e.g., Hoedt et al., 2021; Frame et al., 2023). But any attempt in this direction – including our own – have produced models that are worse than pure machine learning. The section that the reviewer refers to reads "*It would be interesting to show that adding physics to a well-trained ML model adds information*". The emphasis here is on "*adds information"*, i.e. makes the model better. We will clarify this in the revised manuscript.

p.7, l.134ff: I completely agree! There is also no reason not to train ML or any other models with multi- objective, -criteria and/or -variable schemes no mater if long time series are available or not and no mater if large samples are available or not. Any method to (further) constrain the feasible model and/or parameter hyperspace has the potential to help.

> **Response:** While we agree with the comment, it is not what the sentence in line 134 states or suggests. To our understanding it is still an open question as to whether multimodal/multitask training is beneficial for hydrology ML models and we would like to see/produce evidence for that in the future.

p.7, l.136ff: there is similarly plenty of alternative information publicly available for training of models. I do understand that currently most if not all LSTMs are single-variable output models. But is it implausible to think that they can be forced to generate multiple output variables for which data/observations are publicly available either globally (e.g. evaporation, snow cover, storage changes, etc) or in many countries in-situ (e.g. groundwater levels) and that need to be mimicked simultaneously to stream flow? In addition, it would be surprising if LSTMs could not be improved by forcing them to simultaneously reproduce various streamflow signatures (e.g. Flow duration curves, autocorrelation functions, etc) or, what is very effective in PB models, long-term and seasonal runoff coefficients as proxy to enforce at least some level of energy conservation.

Response: No, that is not implausible at all, and we know of at least a handful of groups, us included, that are working on this research direction. Personally, we think this is a very interesting research direction and there are a lot of open questions to explore. If we draw from the general field of machine learning, we should expect an improvement in model robustness/performance, by training on multiple (connected) targets that share underlying processes.

Thank you for this important contribution and I hope you find my thoughts helpful to further strengthen the manuscript! Please note that in the comments above I have added a few references to the work of our group. I have done this for my own convenience and to save time having to search other group's references. Other groups will have produced work that is potentially more suitable to cite here. Please therefore understand these references as mere examples and suggestion and feel under no obligation to use them in any way in you manuscript. Best regards, Markus Hrachowitz

Comments by Anonymous Reviewer in *blue*, our responses in black.

In this study, rainfall-runoff data from 531 watersheds in the camel dataset were simulated utilizing the Long Short Term Memory (LSTM) model. The LSTM model, when trained with multiple basins, exhibited superior performance compared to those trained with a single basin. The authors assert that LSTM streamflow models yield optimal results when trained with extensive and hydrologically diverse datasets, encompassing as many watersheds as feasible. Regardless of the specific objectives a researcher may have for training a machine learning (ML)-based rainfall-runoff model, there appears to be no compelling reason not to employ a large-sample dataset for training purposes. Segregating basins into groups based on hydrologically-relevant characteristics appears to enhance the performance of models trained with limited data, surpassing models trained on randomly grouped basins of comparable size.

I concur with the authors' viewpoints. The research outlined in the manuscript is methodical and enhances our comprehension of the necessity to utilize data from numerous basins for effectively training LSTM models for rainfall-runoff modeling. The statistical analyses conducted by the authors are robust. Nevertheless, significant revisions are imperative prior to considering this manuscript for acceptance.

My suggestions and critiques are enumerated as follows:

Abstract: The abstract contains excessive information regarding the popularity of ML and LSTM in rainfall-runoff modeling. To align with the traditional structure of scientific papers, I propose including more details regarding the methodology implementation, delineating how the LSTM model's performance improves with a larger dataset, and specifically identifying the most crucial influencing factors for LSTM training in rainfall-runoff modeling.

> **Response:** This is a fair point, but we largely do not agree. Scientific papers are typically motivated by an impact statement, and since this paper is a paper about modeling itself (instead of, for example, about how modeling allows us to learn about the real world), the impact statement in this case is about why ML models are important. The analogy would be if we were writing a paper on flood forecasting, we would motivate the paper with a few statements about floods being the most common type of natural disaster, impacting XX people, etc.

Section 1: While it is true that ML algorithms vary, it would be beneficial for the authors to elaborate on why LSTM has gained prominence, particularly due to its development for time series prediction, and elucidate why other ANN systems are seldom employed in hydrological studies.

> **Response:** This is a good point. We will add a sentence or two about this in the revision.

Section 2: Figure 2 presents results that may be surprising to readers. The authors should provide more information on why VIC basin and mHM basin outperform VIC regional and mHM regional. While these findings support the authors' assertions, additional context is required to evaluate their accuracy. Consider transferring relevant information from the appendix to Section 2 for clarification.

> **Response:** To our understanding, it is common knowledge that regionalization strategies under-perform compared to local calibration. Given that this is important to the point of the paper, we will certainly provide some insight into why this is the case.

Section 3: The utilization of 531 CAMEL basins, encompassing more extreme runoff data, to enhance LSTM prediction is comprehensible. However, clarity is lacking regarding Figure 3 and the content between Line 61 and 67. Additionally, it remains unclear why there is a limitation on regional models in Figure 2, and how the bounded vector in LSTM influences runoff prediction as mentioned in Lines 63-64.

> **Response:** Thank you for pointing out sections of the writing that could be improved. This level of detail is sincerely appreciated. We will revise these sections for clarity.

Section 4: Would presenting Figure 5 on a logarithmic scale for the x-axis be more appropriate? This aspect requires further consideration.

> **Response:** We did try a log scale (and a log-log scale) for figures 4 and 5. We felt like the message is clear using the figures in their natural scales, and it is just *slightly* more complicated to glance at a log-scaled figure. Also, this way there is no risk of a reviewer worrying that we were using a figure that exaggerates the real effect.

Section 5: Enhancements to Figure 6 could involve using more contrasting colors to ensure clarity in black-and-white printed versions.

> **Response:** Thank you for noticing this. We will use different line styles in Figure 6 to accommodate printed versions.

Code and Availability: While the code and data are accessible for download, they remain challenging to execute, especially for those unfamiliar with the NeuralHydrology Python package. I recommend providing supplementary materials to furnish more detailed information on these packages.

> **Response:** Unfortunately, this comment is not specific enough to provide us with the information that we would require to add such supplementary material. NeuralHydrology is already as “plug-and-play” as Python packages get. Reproducing our experiments requires only installing the software and running a single command. These steps are well documented in the documentation of NeuralHydrology.

Additionally, NeuralHydrology comes with full documentation and tutorials. You can run the experiments on personal computers (with a single GPU) so that anyone should be able to use it with almost no effort. Additionally, we have previously published a peer-reviewed software paper about this code package (cited in the current manuscript), and we do not see the need to reproduce the contents of that software paper in each publication that uses this codebase.

Considering the significance of the content, it is advisable to integrate essential information from the appendix into Sections 1-5, reserving detailed model settings for supplementary materials.

Response: We believe that the current organization makes the main message of the manuscript more accessible (after all, it is an opinion paper and not a research article). The organization of this paper is designed so that readers can quickly get the main message, and the appendix then also allows people to reproduce some of the experiments that we use to highlight the opinion. To re-emphasize, the purpose of this opinion paper is not to present new research but instead to address a systemic problem in the hydrology literature.

Comments by JD in *blue*, our responses in black.

The concept of summation vs. unit hydrograph

By the title alone of their opinion paper, I admire the frankness of the authors admitting the shortcomings of the LSTM networks when applied to a SINGLE basin.

Response: This is not a correct or meaningful interpretation of the message of this manuscript. The paper does not say that you should not use an LSTM for a single basin, the paper says that if you only want to model a single basin, you should still train the LSTM correctly (i.e., on data from multiple basins). This is not a shortcoming in the sense that we are not illustrating any situation where the model should not be used, we are providing insight into the proper way to use the model.

In the LSTM, the form of the conversion function for the hidden gate h_t, is tanh, e. g., Kratzert et al. (2018, Equation 7). As observed previously by me, this is similar in shape to a summation or S-curve hydrograph in unit hydrograph theory, e.g., Lees (2022, CC1 and CC2 therein).

For the hidden gate, I suggest the authors consider taking one further step of using the first derivative of the conversion function. This is equivalent to using the form of an instantaneous unit hydrograph or impulse response function in convolution integral, e.g., Ding (1974). This, I believe, will inject some hydrologic realism into the LSTM.

The bottom half of Figure 4 for Buffalo Fork, Wyoming (USGS Gage 13011900) for the single LSTM basin model clearly indicates that an impulse response model having a distinct peak time and magnitude characteristic, e.g., Jeong and Kim (2023, Figure 2), would outperform the LSTM as now configured.

A reconfigured LSTM as suggested above may perform as well as, if not better than, the impulse response ones.

Response: The purpose of this manuscript is not to investigate different modifications to the LSTM architecture but rather to make the community aware of the specific training setup that is required when training these (somewhat popular) models. The review has made this same (or similar) comment to previous LSTM publications in open review in HESS-D, and we suggest that if the reviewer is interested in this topic that they could try the experiment themselves. This type of experiment is out-of-scope for the current publication. This could be done, for example, starting from the LSTM implementation in https://github.com/neuralhydrology/neuralhydrology/blob/master/neuralhydrology/modelzoo/customlstm.py.

Comments by JM in *blue*, our responses in black.

The opinion piece draws the attention towards the size and diversity of data used to train LSTM models for streamflow. I'm really happy that the authors put this material together to emphasize the good practice of using large and diverse datasets for models that will reliably predict streamflow at ungauged locations and time periods that are beyond the time periods the models were trained on.

I think the manuscript is well structured and nicely written. The arguments are easy to follow and the figures are appropriate to support the statements. I actually do not have any major comments; the list of minor (easy to address) comments are attached below. I would like to finish with congratulating the authors to this nice manuscript. It was a very enjoyable read and hope it will be published soon.

Best regards,

Juliane Mai

Response: Thank you very much.

Minor:

L 37-38: "It is important to recognize that there is usually no reason in practice to train LSTM streamflow models using data from only a small number of watersheds." —> Couldn't it be that one could run into storage limitations when, for example, wanting to use 10,000s of basins? I know this is not the target audience here, but maybe it should be mentioned that there is an upper limit of how many basins one can use and still be able to train a model with a standard machine.

Response: Even if your dataset becomes too large to be kept in memory (RAM), there are existing alternatives to stream data from disk into memory during model training, which allows training on any dataset that fits on disk.

L 45-52: Section 2: I think it might be helpful to re-iterate that the behaviour the LSTMs show here is the exact opposite compared to the physically-based models (left panel). I'd probably also add panel IDs to the figure (e.g., A and B) and then refer to them in the text. Up to the authors.

Response: Thanks, we will add panel IDs to Figure 2 and we also adapt the first paragraph of Section 2, as suggested by the reviewer.

Figure 2: caption: "for models trained on individual basins (basin) vs. on multiple basins (regional)" —> "for models trained on individual basins (basin; orange coloured lines) vs. on multiple basins (regional; blue coloured lines)"

Response: Thanks, we will apply those suggestions.

L 63: "a vector that is bounded bounded to 1 (-1)" —> "a vector that is bounded to 1 (-1)". Also, what does the "-1" in parenthesis mean? Is the vector bound to the interval [-1,1]? I am guessing the "(-1)" refers to "(limiting)" in the next sentence?! But it is confusing. I'd probably not mention the lower bound because you already made the point clear with the upper limit that can be reached.

Response: Yes, this was sloppy notation. Thank you for pointing this out, we will fix it in the revision. And yes, the LSTM output is bound to the interval (-1, 1) (non-inclusive of the bounds).

L 65: Missing closing parenthesis after "Appendix B".

Response: Thanks.

Figure 3: Wow! This is such an impressive difference!

Response: We agree 🙂

Figure 4 (caption and text): You illustrate the simulation of the single-basin model (13011900). I am assuming that the time period you show is the testing period for the model that was trained with this basin only (no spatial transfer), right? It would be helpful to mention this here as the reader doe not know which 10-year period you use for training vs testing (it's in the appendix I am sure but it would help the flow of reading).

Response: Indeed this is the test period (i.e. neither the train nor validation period), we will make sure to clarify this. We are not really sure if we understand the second part of your comment though. The plot shows simulations from two different models. The top plot shows results from the regional model (i.e. a model trained on the training period of all basins) and the bottom plot shows the results from the single basin model (i.e. a model trained only with the training data of basin 13011900).

Section 3: Caption is missing a question mark at the end.

Response: Thanks, we will fix this.

Section 3: I would finish that section with a clear statement like: "When you train your LSTM model on a single basin, you will likely not be able to predict any peak event in the future." This is what I would expect to read as a response to the question you state in the section title.

Response: Thank you for the suggestion. In general, we appreciate (and like) the idea about adding a summary sentence to the end of each section. We will consider the best way to approach this in the revision.

Section 4: I think it might be confusing when you talk about splits here given that hydrologists usually use "splits" to use one set for calibration and then the remaining for validation. When you say N splits (I think!) you mean that you repeat training/testing N times, right? Like N independent experiments. E.g., "The 531 basins are, for example, split into 5 groups (each around 107 basins). Then the first group of basins (first split) is trained on the training period (Oct 1999-Sep 2000) and then all basins of the first split are tested during validation period (Oct 1980-Sep 1989). After training, the model is evaluated for the ~107 basins of the first split using the testing period (Oct 1989 to Sep 1999). This experiment is then repeated for each of the remaining 4 splits." Not sure if that is entirely correct but it took me quite some reading of the appendix to get to this.

Response: We will try to rephrase parts of this section to make this part easier to understand.

Section 4: Do you think that 531 basins might not be enough because they are not diverse enough? Do you think 531 would be enough when the training period is longer than 10 years? I don't think any additional experiments are required here but it might be nice to have a bit of discussion here. Otherwise it might be really demotivating to know that even >500 basins are not enough.

Response: This is a very hard question that we unfortunately don't have an answer for (yet). First of all, what is "enough"? Enough in the sense of getting good predictions in these 531 basins on a different time split or enough for learning a robust and very general understanding of the different types of catchments included in CAMELS? The CAMELS dataset consists of a diverse set of catchments however the number of basins per type of catchment (e.g. catchment with glacier, catchments in the desert, high mountain catchments) do vary a lot. From our intuition, adding more data (basins and/or longer records) should help learning a better understanding of the underlying processes in these environments. Also thinking about the global context, there are certainly types of catchments that are not included in CAMELS.

L 97 (and caption figure 6): "the curve in Fig. A2" —> There is no curve in figure A2. Do you mean Figure 5?

Response: Indeed, thank you for pointing out this mistake.

Figure 6 caption: For completeness, describe what the dashed orange/green line mean.

Response: We will.

Section 5: I love the description of the experiment and its discussion here. Also the take-home message at the end is great. This is something that Section 3 (take-home) and Section 4 (description experiment) are slightly lacking.

Response: Thanks

Line 128: "good" —> maybe "reliable"?

Response: The problem with the word “reliable” is that we didn’t calculate any reliability metrics. We could say “accurate”, but the scope of the preceding discussion is slightly larger than pure accuracy (e.g., Figure 3). We actually thought fairly carefully about this particular wording in the original writing.

Line 129: “We’ve” —> “We have”

Response: Thanks, will be changed.

Line 131: “Of course, it is trivial (but most likely uninteresting) to beat improperly trained models.” Love that statement!

Response: Thanks.

Figure C1: I think it might be helpful to indicate (on x-axis) which the optimal value is for each metric. Table C1 might also be a good place for it. Up to the authors.

Response: Will be added.

Comments by TN in *blue*, our responses in black.

The manuscript was well written and the topic is very interesting. I really enjoy reading this manuscript. The discussion is thorough, providing a deep understanding of LSTM and good practice. I think there is still room for discussion (related to the questions below). Please feel free to include it if the authors think these points are relevant.

1. *Why is the performance of physical or conceptual models (e.g., mHM and VIC) when calibrated for a single basin ("basin model") better than the one calibrated for multiple basins ("regional model") while that is the opposite with LSTM?*
2. *"Never train an LSTM on a single basin". What makes LSTM so special that we should not train on a single basin? What if we want to model new processes with LSTM and we just have data for a single (or just a few) basin(s)? – we do not have other publicly available data (as for streamflow modeling).*

Response: Related to the first question, the reason for this particular behavior of conceptual models is more-or-less well known (it is related to the "uniqueness of place" problem in hydrology). ML is different because it can learn more general relationships (generally, what this reviewer describes in their comments below), and ML is almost always better (across almost all domains of application) when trained on more data. We will add some of that background to the introduction of the paper.

Regarding the second point, we would like to note that we don't see the LSTM as the swiss-army-knife that should be blindly applied to any problem. One always needs to carefully assess if this specific model is appropriate for the task (and data) at hand. That being said, there are various options that one could think about how DL models can still be used for a task with limited data availability. Probably the most obvious approach to us would be multi-task learning. That is, maybe there is another target with abundant data that is connected to the other task (e.g. discharge and soil moisture). In that case, one could train a model on both tasks together, hoping that learning about task 1 (with abundant data) helps the model to better understand task 2 (with limited data). That all being said, we see this discussion as somewhat off-topic for what we discuss in this manuscript, but hope that our answer helps the reviewer to explore this area.

Blow are possible discussions regarding these questions (just my opinion, please correct me if I am wrong)

Discussion of question 1:

Physically-based (or conceptual) model formulation: The physical processes are conceptualized using mathematical /empirical equations that were based on experimental results or the modeler's understanding of reality. In this sense, PB models already used some kinds of data (because the equations in PB models are based on experimental results

or the modeler's understanding of reality – so equations in PB models are also a kind of data).

PB models: "basin" and "regional" models are often calibrated for a few dozen (or a few hundreds) calibration parameters (I am not sure how many calibration parameters there are for the VIC and mHM that used in this study?). There are uncertainties in the data, model structure, and imprecise description of the physical process via mathematical equations in PB models, the calibration process will adjust calibraiton parameters to compensate for these errors/uncertainties. However, when applying to a large number of basins, the number of parameters might not be large enough to compensate for these errors/uncertainty => leading to a reduce model performance of the "regional" model compared to "basin" model.

ML models: "basin" and "regional" both have a high number of "trainable" parameters (similar to the term calibration parameters in PB) compared to PB models. For example, a basin (regional) model with 16 (256) hidden states, 1 layer of LSTM, and the model head consisting of 1 dense layer could have more than 1,400 (250,000) "trainable" parameters (these numbers were roughly calculated for a native LSTM network in PyTorch). With such models, it can easily overfit the train data and provide poor performance for test data. While the "regional" model with more data the overfitting problem might be avoided (but still possible) and provide a better performance for test data compared to "basin" model. This could be a reason why the performance of the "regional" model is better than the "basin" model with the LSTM network.

Now if we calibrate "basin" and "regional" PB model by adjusting parameters for individual model grid cells (or many groups of grid cells based on certain criteria), which results in a much larger number of calibration parameters – up to the level that the number of calibrated parameters of PB models and the number of trainable parameters of LSTM are the same, will PB behave similar to ML model? (I mean will the performance of the "basin" model is worse than the "regional" model?).

Response: Generally speaking, spatially distributed, grid-based hydrology models are worse than conceptual hydrology models. Our perspective is that hydrological processes are simply not understood well enough to be modeled effectively in explicit ways. There is quite a lot of evidence in hydrology literature that this is the case. An example are the Biosphere-2 hillslope experiments, which measured soil properties in a dense grid along a small (artificial) hillslope, and even with that level of information process-based models could not model saturated flow with meaningful accuracy.

Discussion of question 2:

What makes LSTM so special that we should not train on a single basin? Assume that we have one basin and train the model for daily streamflow simulation for 10 years. In this case, we have 3650 pairs of inputs and outputs– this is already "big data" compared to

other models that use a much lower number data points (e.g., Table 1 – in Zhu et al., 2022; https://doi.org/10.1016/j.eehl.2022.06.001). I would be curious to know if this could be because the number of trainable parameters of the LSTM networks is high or something related to the long-short term memory (that we need more data to "warm-up" the model?).

What if we want to model new processes with LSTM and we just have data for a single (or very few) basin? Could we impose our understanding of the process of interest on the structure of the LSTM model or somehow make LSTM applicable even for a single basin? For example in Figure 4 (page 5), if we know that there are seasonal and trend components, can we model these two components separately and then combine them which might improve the model prediction, etc.. I would very helful it if the authors could give some comments on this.

Response: Please see our response above on our opinion regarding the second question.

Thank you

Tam