Comments by Anonymous Reviewer 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