General comments:

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. 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 (hereafter PB; including the entire spectrum from lumped conceptual to spatially explicit “physics-based” models).

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.

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"

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)

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.

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

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.

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”

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

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?

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 <u>at least</u> 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.

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.

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.

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.

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.

p.7, 132: not sure I fully understand this statement. Did not some recent papers that were partly co-authored 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)?

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 matter if long time series are available or not and no matter if large samples are available or not. Any method to (further) constrain the feasible model and/or parameter hyperspace has the potential to help.

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.

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

References:

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., ... & Rasmussen, R. M. (2015). A unified approach for process-based hydrologic modeling: 1. Modeling concept. Water Resources Research, 51(4), 2498-2514.

Dembélé, M., Hrachowitz, M., Savenije, H. H., Mariéthoz, G., & Schaefli, B. (2020). Improving the predictive skill of a distributed hydrological model by calibration on spatial patterns with multiple satellite data sets. Water resources research, 56(1), e2019WR026085.

Fenicia, F., Kavetski, D., & Savenije, H. H. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. Water Resources Research, 47(11).

Frame, J. M., Kratzert, F., Gupta, H. V., Ullrich, P., & Nearing, G. S. (2023). On strictly enforced mass conservation constraints for modelling the Rainfall-Runoff process. Hydrological Processes, 37(3), e14847.

Gharari, S., Gupta, H. V., Clark, M. P., Hrachowitz, M., Fenicia, F., Matgen, P., & Savenije, H. H. (2021). Understanding the information content in the hierarchy of model development decisions: Learning from data. Water Resources Research, 57(6), e2020WR027948.

Gupta, H. V., Sorooshian, S., & Yapo, P. O. (1998). Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. Water Resources Research, 34(4), 751-763.

Gupta, H. V., Wagener, T., & Liu, Y. (2008). Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydrological Processes: An International Journal, 22(18), 3802-3813.

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., & Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. Water Resources Research, 48(8).

Hoedt, P. J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G. S., ... & Klambauer, G. (2021). Mc-lstm: Mass-conserving lstm. In International conference on machine learning (pp. 4275-4286). PMLR.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., ... & Gascuel-Odoux, C. (2014). Process consistency in models: The importance of system signatures, expert knowledge, and process complexity. Water resources research, 50(9), 7445-7469.

Hrachowitz, M., Stockinger, M., Coenders-Gerrits, M., van der Ent, R., Bogena, H., Lücke, A., & Stumpp, C. (2021). Reduction of vegetation-accessible water storage capacity after deforestation affects catchment travel time distributions and increases young water fractions in a headwater catchment. Hydrology and Earth System Sciences, 25(9), 4887-4915.

Hulsman, P., Savenije, H. H., & Hrachowitz, M. (2021a). Learning from satellite observations: increased understanding of catchment processes through stepwise model improvement. Hydrology and Earth System Sciences, 25(2), 957-982.

Hulsman, P., Hrachowitz, M., & Savenije, H. H. (2021b). Improving the representation of long-term storage variations with conceptual hydrological models in data-scarce regions. Water Resources Research, 57(4), e2020WR028837.

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. Water resources research, 42(3).

Leavesley, G. H., et al. "The modular modeling system (MMS)—The physical process modeling component of a database-centered decision support system for water and power management." Water, Air, & Soil Pollution 90 (1996): 303-311.

Mendoza, P. A., Clark, M. P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G., & Gupta, H. (2015). Are we unnecessarily constraining the agility of complex process-based models?. Water Resources Research, 51(1), 716-728.

Nijzink, R. C., Almeida, S., Pechlivanidis, I. G., Capell, R., Gustafssons, D., Arheimer, B., ... & Hrachowitz, M. (2018). Constraining conceptual hydrological models with multiple information sources. Water Resources Research, 54(10), 8332-8362.

Pokharel, S., Roy, T., & Admiraal, D. (2023). Effects of mass balance, energy balance, and storage-discharge constraints on LSTM for streamflow prediction. Environmental Modelling & Software, 166, 105730.

Samaniego, L., Kumar, R., & Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. Water Resources Research, 46(5).

Tempel, N., Bouaziz, L., Taormina, R., van Noppen, E., Stam, J., Sprokkereef, E. & Hrachowitz, M. (2024). Vegetation Response to Climatic Variability: Implications for Root Zone Storage and Streamflow Predictions. Hydrology and Earth System Sciences Discussions, EGUsphere [preprint]