# Review comments (v2) for “Technical note: Does Multiple Basin Training Strategy Guarantee Superior Machine Learning Performance for Streamflow Predictions in Gaged Basins?” by Tran et al.

Author: Frederik Kratzert

This is my second solicited review of this manuscript. The first review was for a different journal where the manuscript was rejected in the first round. Therefore, I never had the chance to see the replies/thoughts of the authors to my comments.

Compared to the version I previously reviewed, most of the manuscript remains unchanged. And while in the previous review I wasn’t sure if much of the manuscript was based on a misunderstanding, as well as mistakes in the data analysis that led the authors to wrong conclusions, I now have to believe that the authors disagree with my comments and stand behind this manuscript.  And maybe because of that, I struggled much more with this review than previously.

Looking at the HESS manuscript type definitions, I think paper does not qualify as a technical note, but much rather could be seen as a direct reply to our recently published HESS Opinions paper, called “HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin” (Kratzert et al, 2024). The core part of that opinion paper can be summarized in two short points:

- If you have multiple basins with streamflow observations and the same kind of input data, you are on average (much) better training a multi-basin than training single basin models.
- LSTMs suffer from a saturation problem, which is especially pronounced in single basin models and to some degree alleviated (for most basins) when training a multi-basin model.

From reading this discussion paper, my feeling is that the authors took the (on purpose provocative) title of our opinion paper a little bit too seriously, especially the word “*Never”* otherwise I have no explanation for this manuscript. So if the sole purpose of the authors is to answer the question in their title, then the answer is “No”. This claim was also never made and the question was already answered in the opinion paper that the authors refer to, which however they seem to have missed. For more details on this particular point, see Section 1 of my review.

Unfortunately this is not the only false accusation about conclusions / statements we apparently have made and not the only point where the authors ignored entire sections of our manuscript that do not align with their story.

The most critical point however is the model comparison. Since it seems like the authors try to not only show that ML does not *guarantee superior performance*, again, a claim that was never made, but only that training single basin models is generally not as bad as presented in our opinion paper, the setup of the model comparison is critical.

- The authors try to derive general recommendations from a comparison of models trained with different data qualities (local/global, high/coarse resolution, reanalysis data / real-time data, simulation model vs. operational flood forecasting model), for details see Sect. 2 of my review.
- In another comparison, where some of the models were trained by the author themselves, they effectively compare their models in a *gauged* setting to a multi-basin model in an *ungauged* setting, essentially showing that the multi-basin model is *as good or better ungauged* than their single basin models *gauged.* For details see Sect. 4 of my review.

I am not sure which way I see forward for this manuscript. In any case, the model comparison needs to be corrected to be able to have a scientific debate about “best practices” (L 61).

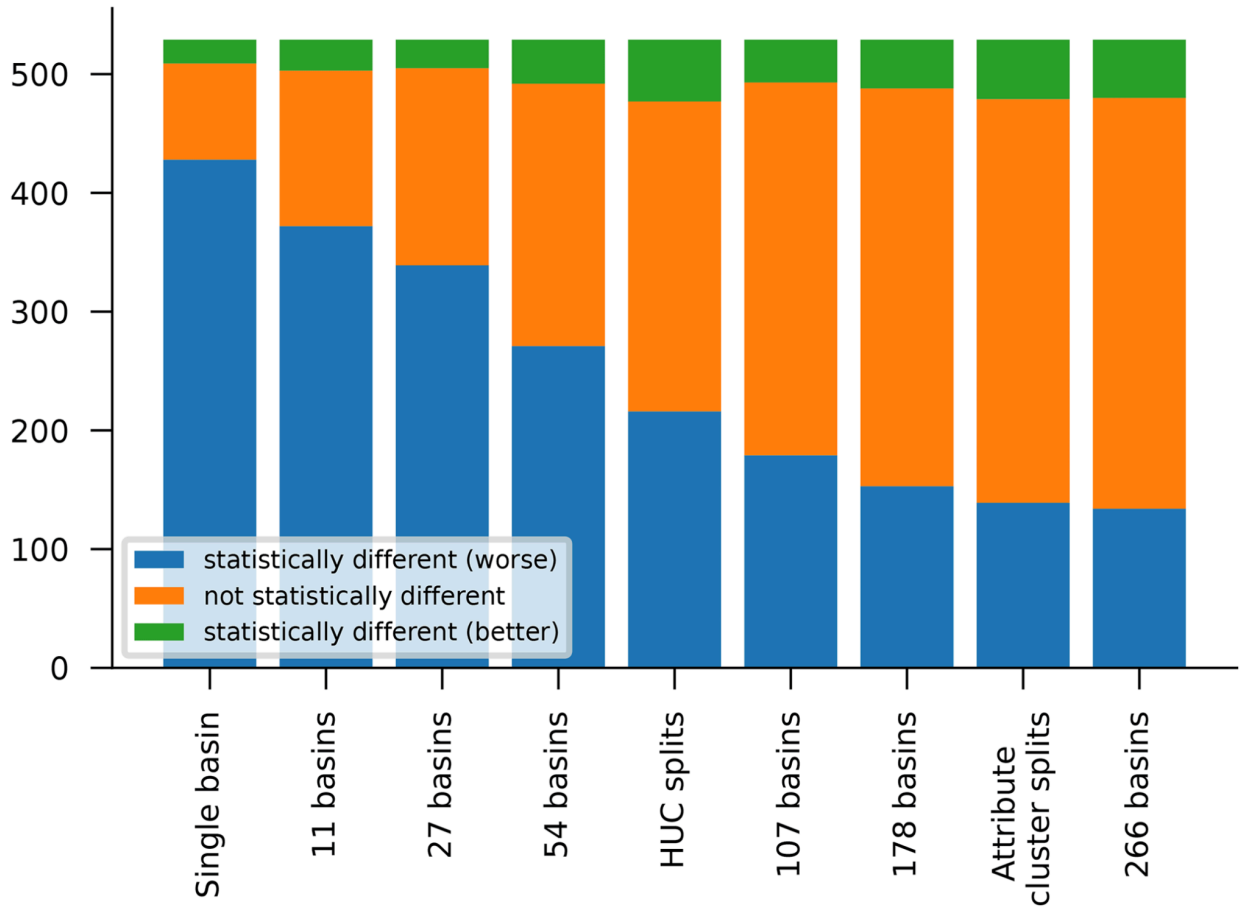
## 1. Does Multiple Basin Training Guarantee Superior Performance?

The title of this manuscript raises this question which according to the authors is a conclusion of our opinion paper. However, this is wrong, to a degree that I almost wonder if the authors only read the title. Kratzert et al. (2024) have an entire section dedicated to this topic, called “*Are bigger models better everywhere?*” (Sect. 6), stating in the very first sentence that:

“*Even though the best model, on average, is the model trained on all 531 CAMELS basins, it is not the case that the model trained on all 531 CAMELS basins is better in every basin*”

Further there is a dedicated analysis on this topic including Fig 7 (copied here for convenience) that shows that there are a number of basins, where the single basin model outperformed the regional model with statistical significance (first column) and that also smaller multi-basin models have a couple of basins where they are better than the larger model. So to answer the title of this manuscript: No, bigger models do not guarantee superior performance. But as said above, this was also never a claim that was being made. However, to cite further from Kratzert et al. (2024)

*“All models perform worse than the full regional model in more basins than they perform better in.”*




And further,

"*We have not found a way to (reliably) predict which model will perform best in any particular basin. It is not possible to use metrics from the training period or validation period to (reliably) choose the best model in the test period. Additionally, we have tried extensively to construct a separate predictor model that uses catchment attributes and/or hydrological signatures to predict whether one model will perform better or worse than other models in specific basins. We have not been able to construct a model that performs well at this task. Details of these predictability experiments are out of the scope of this paper, but a relevant example was given by Nearing et al. (2024).*"

To conclude: I struggle to understand how the authors thought that anybody ever made the claim that "superior performance" can be guaranteed when training multi-basin models vs. single basin models.

## 2. Model comparison

One main concern I have with this manuscript is the model comparison. I already noted that in my previous review but the authors have largely ignored my comments, as this part of the manuscript is mostly unchanged.

In science, if we want to investigate if A is better than B, then we have to do this in a setting where we can exclude (as much as possible) any other factor that impacts the results of this experiment.

This is not what has happened here. The authors picked simulations from different studies that use different data and train models for different purposes to then base their conclusion on this model comparison.

Taking one step back, the main source for uncertainty/error in the rainfall-runoff models come from the quality of the input data, especially (but not only) the weather input data. Here however, the authors took model simulations from different models in different studies being trained on different data (with different quality and temporal availability, i.e. real-time available or reanalysis) to support the conclusion of their manuscript.

- Model G (Nearing et al. 2024), an operational flood forecasting model, relies on *globally available* forcing data that is *available in real-time*. For the hindcast data, this is CPC precipitation, IMERG precipitation, as well as data taken from historic ECMWF IFS-Forcasts and GraphCast.
- Model R, three different regional LSTMs trained with *high resolution / high quality reanalysis data*, which are only locally available. The different model R's are from
    - Kratzert et al. (2024) for the US using Daymet + Maurer + NLDAS and being trained in a gauged setting.
    - Kratzert et al. (2019) for the US using NLDAS and being trained in an *ungauged* setting.
    - Lees et al. (2021) for Great Britain using CEH-GEAR + CHESS-PE + CHESS-met being trained in a gauged setting. The authors state in L116 that the results from Lees et al. (2021) are for *ungauged* basins, which is wrong. The study by Lees et al. only includes *gauged* experiments.
- Model S from Kratzert et al. 2024, are single basin models being trained Daymet + Maurer + NLDAS (i.e. same as above).

- S1-S6, the only set of models the authors train themselves, using ERA5-Land data and in the case of S-6, even lagged streamflow data as input.

It makes absolutely no sense comparing models forced by different categories of data (in terms of real-time/reanalysis, coarse/fine resolution, global/local) to make a general statement about the best model training setup. And taking it to the extreme, it makes even less sense to compare a global operational flood forecasting model with a focus on ungauged regions to a

single basin model that gets lagged streamflow as input (S-6). Did anyone really think that a global operational flood forecasting model that is purely based on coarse resolution weather data is better than a model that sees yesterday's streamflow as input?

The point being raised by Kratzert et al. (2024) is: If you have multiple basins with the same kind of data, then it is better to train a multi-basin model than individual single basin models. Obviously, using lagged streamflow as input makes even a single basin model better than a global, pure simulation model. The question is: Is a multi-basin model with lagged streamflow as input better on average better than a single basin model with lagged streamflow? Similarly, it is rather obvious that a regional model with regionally available, high resolution reanalysis data is better than another model that was trained on globally (and in real-time) available weather data.

If the authors want to show anything else than that you can find a setting where single basin models are better than multi-basin models, again, something that was already shown in Kratzert et al. (2024) and never claimed to be otherwise, then my recommendation, as in my previous review is the following:

The entire model comparison needs to be done in a controlled environment where all models, multi-basin (be it regional or global) and single basin models, have the same data available. Fun fact: This is exactly the experiment that was done in Kratzert et al. (2024) and the results are known by now.

## 3. Sect. 3 “Previous Research Using Single-Basin Trained ML: Was It a Mistake?”

This entire section is dedicated to the results of the literature review we presented in Sect. 1 of our opinion paper. I left basically the exact same comments in my previous review and they were essentially ignored in this resubmission. Here are a few points the authors make that I think are wrong:

- Having read the studies in question myself, I know that a lot of them (IIRC the majority) use historic streamflow as input. Getting an NSE above 0.75 for an autoregressive model is not a sign of an excellent model, much rather it is relatively simple to achieve, given the high autocorrelation of streamflow in time. For that reason, ad hoc “adequacy” thresholds like NSE above 0.5 or above 0.75 are not relevant for autoregressive models.
- A statement like “*the single basin model is above NSE 0.75”* does not tell you if a different approach (e.g. regional LSTM) would not be significantly better.
- And more general: When performing benchmarking studies and making general statements like the one being done here, judging by a single metric is not really the way forward.

- L 106 “*studies have successfully passed through peer review processes.*” is certainly no criteria for “*no evident flaws in model training approaches*”. I don’t know who needs to hear this, but this is such a wrong statement that I don’t even want to expand on this.

## 4. Sect. 4 “Does Multiple-Basin Training Consistently Outperform Single-Basin Approaches? Insights from Experimental”

This Section has changed to some degree from the previous version of this manuscript. Two things changed:

- The authors trained their single basin models (S1-S6) using Caravan data, which in this case consists of ERA5-Land data.
- The authors corrected an one-off error in their metric computation, which made them underestimate the performance of model G in all evaluations.

However, the following points remain:

- The simulations of model G from Nearing et al. (2024) are from a k-fold cross validation experiment. This effectively means that all predictions for all gauges are from an *ungauged* setting.
- On the other hand, model S1-S6 are naturally models trained in a *gauged* setting.

On top of that, but this only has a minor impact, models S1-S6 use ERA5-Land as input, a forcing product that is not available in real-time and includes data assimilation, a forcing product that is not being used by model G in the operational setting.

**So what the authors effectively show in this section is that model G is as good, or even better (see Table B.2), *ungauged* than all of their single basin models that are not relying on lagged streamflow as input *gauged.* I think this is not what the authors wanted to show but it is actually a remarkable result.**

## 5. More equals better!?

A point that the authors seem to suggest that we made in our opinion paper is that “more data equals better models” and therefore that model G, a global model being trained on thousands of gauges should be better than any of the other models, being trained on just a few hundreds of basins. While the results in Kratzert et al. (2024) are limited to the CAMELS dataset, we have an entire section (Sect. 5 “*Is hydrological diversity always an asset?*”) dedicated to this question. This section contains the following sentences:

“*Figure 6 provides evidence that there might be ways to construct training sets that could potentially result in better models than simply training on all available streamflow data. This*

*conclusion is hypothetical because in all the examples shown in Fig. 6, models trained on any subset of the 531 CAMELS basins performed worse, on average, than models trained on all 531 CAMELS basins. However, separating the training set into hydrologically similar groups of basins results in models that perform better than models trained on random basin groups of similar size.*"

So as in some of the points above, I struggle to see where this claim was supposedly made and why the authors don't include this section of our opinion paper into their discussion.

## Line by line comments

- L28: Funnily enough Kratzert et al. (2018) is probably not the paper you want to cite here, as the LSTM model in that paper was worse than the SAC-SMA in most of the settings. The correct paper to cite here would be Kratzert et al. (2019).
- L 31ff: Most of the points in this paragraph are not any different between ML models and PB models.
    - Where ML people "fine-tune ML architectures and hyperparameters" PB people refine the process implementations in their models and which processes to include.
    - Everybody who works with models constantly "explores alternative training approaches".

- L 51: "*there is no comparison of their model trained using data from multiple basins versus individually trained, basic(sic!)-specific models*", when referring to Kratzert et al. (2024). What do you mean here? The entire opinion is a comparison of a multi-basin model vs individually trained, basin specific models. You even use both types of modeling approaches from that paper in your model comparison.
- L61: There are certainly "best practices". Your title raises the question if multi-basin training can *guarantee* superior performance, and it can't. But if we speak about "best practices" in the average case, then yes, they exist and where shown more than once and Kratzert et al. (2024) is just one example. If you really want to contend the established best practices, then you will have to do model comparison in a controlled setting as described above.
- L68ff: I think a more nuanced view should be considered here.
    - a) Taking the most recent paper from the list of references (Addor et al. 2020), which by now is 5 years old, a significant amount of new data has been made publicly available, covering thousands of stations in tens of countries. This trend, to some degree, has been fueled by the increasing amount of large sample hydrology studies, not exclusively but for sure also including ML applications. While some regions remain white spots on the map, I struggle to see how large-scale ML applications have a detrimental effect on the publication of data,

while I can certainly see how single basin applications would have such a detrimental effect.
    - b) If your point is that single-basin approaches have a benefit in regions with only point-based meteorological observations, then you should go ahead and show this, instead of using CAMELS-GB and CAMELS (US).
- L86 Using Knoben et al. (2019) to justify performing model comparison purely based on the Nash-Sutcliff Efficiency almost comical. The last sentence of their paper reads as follows "*More generally, a strong case can be made for moving away from ad hoc use of aggregated efficiency metrics and towards a framework based on purpose-dependent evaluation metrics and benchmarks that allows for more robust model adequacy assessment.*"
- L89 Remove quotation marks around mistakes.
- L89f: Is this list a citation from someone or your own thoughts? If these are references to findings by others, please cite the relevant papers. If they are your own thoughts, please extend further on what you base these statements on. Furthermore, are these unique to ML models? I think all of these points hold for any kind of model and are not specific to ML. Point 1 though is not a mistake I would say, but rather a problem? Because what is the mistake if the trained/calibrated model has poor performance? It is a fact and (hopefully) there is a reason behind it that could be changed to get a better model. But having a model with poor performance is not a *mistake.*
- L 107f: This concluding sentence is wrong on so many levels. I commented above your take of the literature review and certainly what you show here is not contradicting the claim that single-basin training strategy is generally the wrong thing to do.
- L116 as well as Fig 1: The paper by Thomas Lees et al. does not include results for ungauged basins.

## References

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., and Mendoza, P. A.: Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges, Hydrological Sciences Journal, 65, 712-725, 2020.

Knoben, W. J., Freer, J. E., and Woods, R. A.: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrology and Earth System Sciences, 23, 4323-4331, 2019

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using long short-term memory (LSTM) networks, Hydrology and Earth System Sciences, 22, 6005-6022, 2018.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrology and Earth System Sciences, 23, 5089-5110, 2019.

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin, Hydrology and Earth System Sciences, 28, 4187-4201, 2024.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., and Dadson, S. J.: Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models, Hydrology and Earth System Sciences, 25, 5517-5534, 2021

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., and Metzger, A.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559-563, 2024