Dear HESS Editor,

Thank you for handling the manuscript. R1 asked why we didn't use Daymet data --- we have included results from the Daymet data for comparison. R2 asked us to show regional breakdown of model comparisons which we have provided. Other than that, reviewers have mostly sought clarifications like code availability and smaller modifications. We have thoroughly revised the manuscript, and believed the manuscript is now in a suitable state for HESS. In this file we use *italic red text* showing the *reviewers' comments* and black text showing our responses. The line numbers inside the brackets refer to the line number of the revised manuscript (without track change).

Best
Chaopeng

RC1: 'Comment on hess-2022-245', Anonymous Referee #1, 24 Oct 2022

The paper proposes a novel hydrologic modeling approach taking advantage of the recent deep-learning techniques. It appears this approach is quite useful for ungaged basins. It is well written. The source code, however, is not very well structured and the example provided is not easy to follow. I have asked my ph.D. student who has multi-year's experience in Python to run the source code and repeat the example (https://zenodo.org/record/7147450), but he did not succeed. The error message he got is below. I'd suggest that the authors perform careful check themselves, and also ask a third-party to independently verify their code and example. I will ask my student to give it another try later on.

==========

Traceback (most recent call last):

File "dPLHBVrelease/hydroDL-dev/example/dPLHBV/traindPLHBV.py", line 96, in

camels.initcamels(

File "dPLHBVrelease/hydroDL-dev/example/dPLHBV/../../hydroDL/data/camels.py", line 549, in initcamels

calStatAll()

File "dPLHBVrelease/hydroDL-dev/example/dPLHBV/../../hydroDL/data/camels.py", line 388, in calStatAll

x = readForcing(idLst, forcingLst)

TypeError: readForcing() missing 2 required positional arguments: 'fordata' and 'nt'

Citation: https://doi.org/10.5194/hess-2022-245-RC1

AC2: 'Reply on RC1', Chaopeng Shen, 24 Oct 2022

Sorry about this. It just came to our attention that there was a bug for a fresh download. We updated the code release and at the Zedono release you should see the following information about this bug fix. Could you please give it a try again?

==========

This release contains the codes and related data to train models with differentiable parameter learning (dPL) applied to HBV backbone as shown in these papers below. Please read the **instruction file** and this **BUG FIX FILE (**https://bit.ly/3TOKmqK**)** for running the released codes.

Citation: https://doi.org/10.5194/hess-2022-245-AC2

RC2: 'Reply on AC4', Anonymous Referee #1, 24 Oct 2022

My student gave it another try, and it worked this time.

Congratulation to the authors on the excellent work!

Citation: https://doi.org/10.5194/hess-2022-245-RC2

We have updated our code release to a new version with some small issues fixed at this zenodo release https://doi.org/10.5281/zenodo.7091334 . We also created a detailed instruction file and a help file (https://bit.ly/3TOKmqK) logging some common issues to help readers run and use our differentiable models.

RC3: 'Comment on hess-2022-245', Anonymous Referee #2, 11 Nov 2022

This study analyzes the ability of deep learning, and physics-informed learning models to make predictions in regions that are outside of the training set. This is an interesting problem, particularly in testing the limits of learning-based models to make predictions in conditions that are outside of the training set. This paper is a valuable contribution to the hydrological modeling literature, and would like to see it published. While there are some wording issues (listed below), and some issues (also listed below) that I would like to see addressed. In particular, there are a couple of points on the training procedure for the LSTM, which limit its performance, ensembling randomly initialized models and including multiple precipitation products are known to boost the LSTM performance in other experiments, it would be good to check if that would

have the same result for ungauged regions. In terms of presenting results, since this paper is about region specific (as in regions held out of the training set) models, I would hope to see region specific results, which are mentioned in the text (in terms of model parameters), but results are not quantified nor plotted.

Thanks a lot for the to-the-point suggestions. We replied in detail to each comment as below.

Line 18: Is "PUR" an acronym for "Prediction in Ungauged Region", or is it a general term for "regionally held out basins"?

PUR is an acronym for "Prediction in Ungauged Regions" which means holding out large continuous regions for spatial extrapolation tests as proposed in Feng et al., 2021. We have modified the original statement as *"For prediction in ungauged regions (PUR, regional holdout test representing spatial extrapolation in a highly data-sparse scenario)"{line 19}* to avoid confusion.

Feng, D., Lawson, K., & Shen, C. (2021). Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data. Geophysical Research Letters, 48(14), e2021GL092999.

Lines 148-149: Can you please clarify the training periods for all the models. You mention that "Each training instance had two years' worth of meteorological forcings, but the first year was used as a warmup period so the loss was only calculated on the subsequent one year of simulation", this reads to me that your models were trained on just one year of data. This isn't the case, is it? I imaging that, in the training process you cycle through many more years, you just train the model with batches of these individual year? Oh. I read on line 245 that "we used only 10 years of training period". Okay, can you maybe re-word this?

Thanks for reminding us of this potential confusion. We want to clarify that the training period and the length of training instances are two different concepts. Training period refers to the time periods over which data was made available for training, but each iteration, we could just take smaller sequences from the training data, form a minibatch, run the update and complete a weight update. This is frequently done in big data training. More details are as follows.

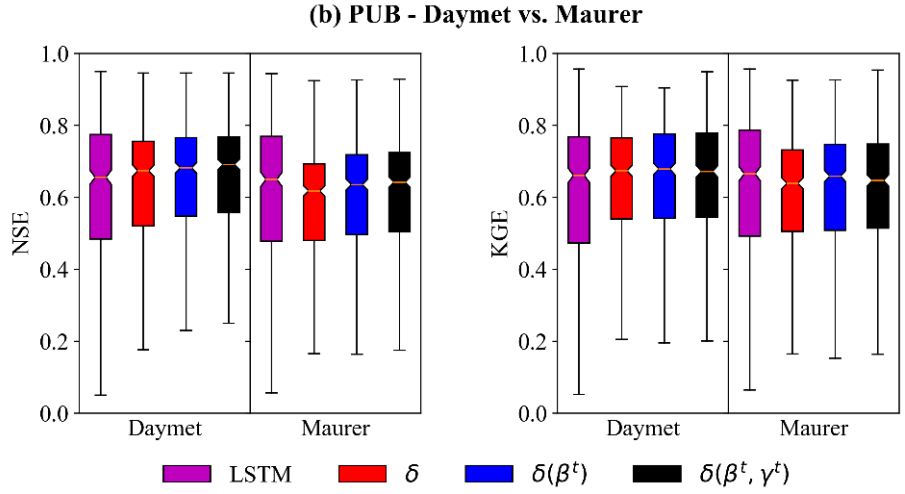
For the training period of spatial generalization tests (PUB and PUR, training and testing in different basins), we originally stated in line 190 and now modified it slightly as *"The study period was from October 1, 1989, to September 30, 1999. These spatial generalization tests were trained and tested in the same time period but for different basins" {Line 212}.* We use data from this ten-year period to train the models. As for the statement in line 148 as "Each training instance had two years' worth of meteorological forcings…", this is talking about how

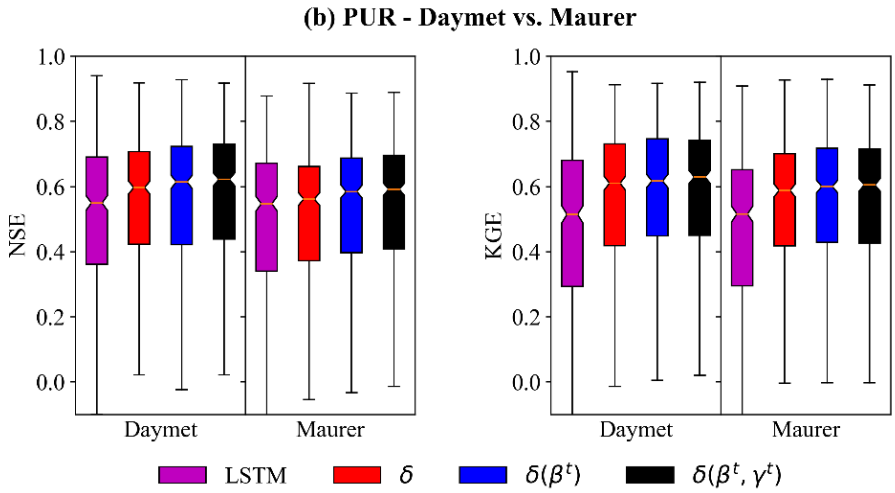
we form a minibatch consisting of short training instances from the whole 10-year period. We added the following clarifications:

"Deep learning models need to be trained on minibatches, which are collections of training instances running through the model in parallel, to be followed by a parameter update operation. In our case, a minibatch is composed of 100 training instances, each of which contains two consecutive years' worth of meteorological forcings randomly selected from the whole training period for one basin. The first year was used as a warmup period, so the loss was only calculated on the second year of simulation. The model ran on this minibatch and the errors were calculated as a loss value, and then an update of the weights was applied using gradient descent." {Line 159}

Line 169: Why was Maurer selected? Especially since many studies suggest Daymet is the more informative forcing, including Feng et al., 2022? It is also the case that using a combination of the three forcing products from CAMELS results in improved model performance (Kratzert et al., 2021), can you expand on your decision in the context of using multiple precipitation sources?

This is a good point. Actually, we hesitated for a while regarding which forcing to use for showing our results. We finally chose Maurer because we want to compare with traditional regionalized models like MPR+mHM , which used that forcing. We agree that different forcings will impact the modeling results and the key point is the used forcing should be comparable for performance comparison. In the revised manuscript, now we also show the performance under Daymet forcings, and compare the results to using Maurer forcings. In the previous manuscript, LSTM was slightly better for the PUB (interpolation under data-sparse scenario) case, but, in the revised manuscript and with Daymet forcing, the differentiable models are better than LSTM for PUB. We added these below two figures as new Figure 3b and 6b, and modified related discussions in the main text (copied below).





"For the randomized PUB test, which represents spatial interpolation in a data-dense scenario (Figure 1a), the δ models approached (under the Maurer forcings) or surpassed (under the Daymet forcings) the performance of the LSTM on the daily hydrograph metrics. Under the Maurer forcing, δ(βt,γt) had a median PUB NSE of 0.64, only slightly lower than LSTM (0.65) and considerably higher than MPR+mHM (0.53, this model is in sample -- all basins were included in training). When one moves from in-sample prediction to PUB, the performance of all types of models drop, as demonstrated by δ(βt,γt) (Figure 3a). For KGE, δ(βt) and δ(βt,γt) models had median values of 0.66 and 0.65, respectively, which were essentially the same as LSTM, but also had a smaller spread (Figure 3a). LSTM had lower errors for FLV and FHV than the δ models (Figure 3a), which is possibly because LSTM is not subject to physical constraints like mass balances and therefore possesses more flexibility in terms of base and peak flow generation than HBV." {line 235}

"Under the Daymet forcings, δ(βt) and δ(βt,γt) models reached NSE(KGE) median values of 0.68(0.68) and 0.69(0.67), respectively, surprisingly higher than LSTM at 0.66(0.66) (Figure 3b). Both the LSTM and δ models showed better performance when driven by Daymet forcings, which is consistent with previous studies using different forcings (Feng et al., 2022; Kratzert et al., 2020a), but δ models improved even more noticeably, showing a clear outperformance of the other models. This result suggests that precipitation in the Maurer forcing data may have a larger bias and, as δ models conserve mass and cannot by default apply corrections to the precipitation amounts, they are more heavily impacted by such bias." {line 245}

"Under Maurer forcings, the median NSE values for LSTM, δ, and δ(βt,γt) models were 0.55, 0.56 and 0.59, respectively, and the corresponding KGE values were 0.52, 0.59 and 0.61, respectively. The performance gap between LSTM and δ models was larger under Daymet forcings. The LSTM had a minor performance gain when using Daymet forcings, while the δ models had significant performance improvements. The median NSE (KGE) values for LSTM, δ, and δ(βt,γt) models were 0.55(0.51), 0.60(0.61) and 0.62(0.63), respectively. We see that for the low flow dynamics, δ(βt) had a slightly smaller low flow bias (FLV). For high flow, δ models still had negative biases but they were smaller than those of LSTM (Figure 6a)." {line 344}

We agree that using multiple forcings can be beneficial when pursuing the best performance for DL models, and DL models also have advantages to efficiently integrate multi-source data compared with physical models. Using multiple forcings will require more effort for differentiable models which maintain mass balance, and we will get to it in the future as our method is so new. We added some discussions on this as copied below from the revised manuscript. However, the key point here is to make all experiments comparable (keep the same setup) rather than gain the best performance. It's worth emphasizing that we don't aim at invalidating the power of LSTM models, and LSTM is part of the parameterization pipeline for the differentiable models. Instead, we want to show how well the differentiable models can generalize in space with the previous regionalized methods providing context. We also added discussions in the revised manuscript to further clarify this point.

"This study demonstrated how well the novel differentiable models can generalize in space with other regionalized methods providing context. To ensure comparability across different models, we have chosen the same setups, e.g., meteorological forcings, training and testing samples and periods, and random seeds, rather than configurations that would maximize performance metrics. This work also does not invalidate deep learning models as valuable tools, as LSTM is a critical part of the parameterization pipeline for the differentiable models. The point of differentiable models is to maximally leverage the best attributes of both deep networks (learning capability) and physical models (interpretability). Several strategies can be applied to enhance the pure data-driven LSTM's performance as shown in earlier studies. For example, some auxiliary information like soil moisture can be integrated by a kernel to constrain and enhance the extrapolation (Feng et al., 2021). LSTM models can utilize multiple precipitation inputs simultaneously to gain better performance (Kratzert et al., 2020b), which can be more complicated to achieve for models with physical structures. Ensemble average prediction from different initializations (Kratzert et al., 2020b) or different input options (Feng et al., 2021; Rahmani et al., 2021) can often lead to higher performance metrics. Here, however, we used a less computationally-expensive but comparable setup without these strategies applied, which can certainly be studied in the future." {line 474}

There should be a direct link to the analysis done for this paper. I browsed around the HydroDL github repository, and it was not clear to me where I should look for the code that was used for these particular experiments. NEVERMIND ABOUT THIS. I NOW SEE THE AUTHOR'S RESONSE TO ANOTHER REVIEWER.

Yes. For anyone who has not seen it, the codes were released at zenodo with this link to access: https://doi.org/10.5281/zenodo.7091334

The issue of ungauged regions is not particularly relevant to the United States (U.S.), but I do see the value of using the U.S. gauged basins for this experiment. Other groups (Le et al.,

2022) have done ungauged region experiments outside the U.S., and this could be a bit more compelling. Perhaps this is worth some discussion in the paper?

We agree that PUR is very important for global hydrologic modeling but the significant uncertainty involved in spatial extrapolation has not gained enough attention. There should be extensive work on global hydrologic modeling using differentiable models in the future. The final aim would be better modeling the global ungauged regions and we think the most difficult problem is the cross-continent prediction. We added the following discussion about the global modeling topic to motivate future research for PUR at the global scale.

"We used CONUS basins and large regional hold-outs to examine the spatial generalization of different models. PUR is a global issue because many large regions in the world lack consistent streamflow data. We ran experiments over the CONUS in this paper to ensure comparability with previous work and to benchmark on a well-understood dataset. It has been demonstrated that models trained on data-rich continents can be migrated to data-poor continents: Ma et al. (2021) showed that deep learning models may learn generic hydrologic information from data-rich continents and leverage the information to improve predictions in data-poor continents with transfer learning. More recently, Le et al. (2022) examined PUR in global basins for monthly prediction with traditional machine learning methods, and the results demonstrated the difficulties of this issue. In future work, we will establish differentiable models for a large sample of global basins by integrating modern DL and physical representations that have shown promising spatial generalizability, and examine their value for accurate daily PUR at the global scale." {Line 487}

Line 245: You mention that there was no ensembling of models trained from random initialization. But then go on to say that you used the same settings as Kratzert et al., 2019, but they used ensembles of 10 models trained with random initializations. From their paper: "Because of this, the LSTM-type models give better predictions when used as an ensemble. It is not necessarily the case that if one particular LSTM model performs poorly in one catchment that a different LSTM trained one exactly the same data will also perform poorly." This is generally an accepted practice when using deep learning models. Can you explain further why you decided not to use model ensembles?

Thanks for pointing out ensemble modeling. As we stated in a previous response, our primary goal is to make the setup of all models and experiments comparable to examine the spatial extrapolation capabilities. With our LSTM, we already repeated the same setup (best performance setup) as the ensemble LSTM job in Kratzert et al., 2019 for benchmarking and found identical results (Feng et al., 2021). So we already know that our LSTM can be used as a good benchmark level.

When comparing different models for spatial generalization, we employed a less expensive but consistent setup without using the ensemble simulations. We also kept all the modes using the same random seed for initialization. The PUR and PUB tests already use cross validation with

7 and 10 models running for each experiment which can ensemble will be too time-consuming. In addition, we wish to do a more thorough evaluation of various ensemble methods in the future, e.g., we can even make an ensemble by perturbing the inputs to neural networks (referred to as an input-selection ensemble) as shown in Feng et al., 2021. It is indeed possible that a particular ensemble method would give advantages to certain methods under different conditions, but we think this complication would need a lot more computational and experimental time to examine, and is outside the scope of this work. We discussed this point as a limitation and a future work plan item and we hope the reviewer can see that there are many things to do on our plate right now with the new differentiable models - their optimal setup is still very much fluid.

In the caption of Figure 3 we gave explanations as *"The PUB was run in a computationally economic manner to be comparable to other models while also reducing computational demand: we used only 10 years of training period, did not use an ensemble, and used a lower k-fold. When we previously ran the experiments using the same settings as Kratzert et al. (2019), our LSTM was able to match the PUB performance in their work (Feng et al., 2021)."{line 274}*

We added this to the Discussion: *" Several strategies can be applied to enhance the LSTM's performance as shown in earlier studies. For example, some auxiliary information like soil moisture can be integrated by a kernel to constrain and enhance the extrapolation (Feng et al., 2021). LSTM models can utilize multiple precipitation inputs simultaneously to gain better performance (Kratzert et al., 2020), which can be more complicated to achieve for models with physical structures. Ensemble average prediction from different initializations (Kratzert et al., 2020) or different input options (Feng et al., 2021; Rahmani et al., 2021) can often lead to higher performance metrics. Here, however, we used a less computationally-expensive but comparable setup without these strategies applied, which can certainly be added in the future." {line 479}*

Please also kindly refer to the responses we added above (on page 6 of this Response) for the added paragraph.

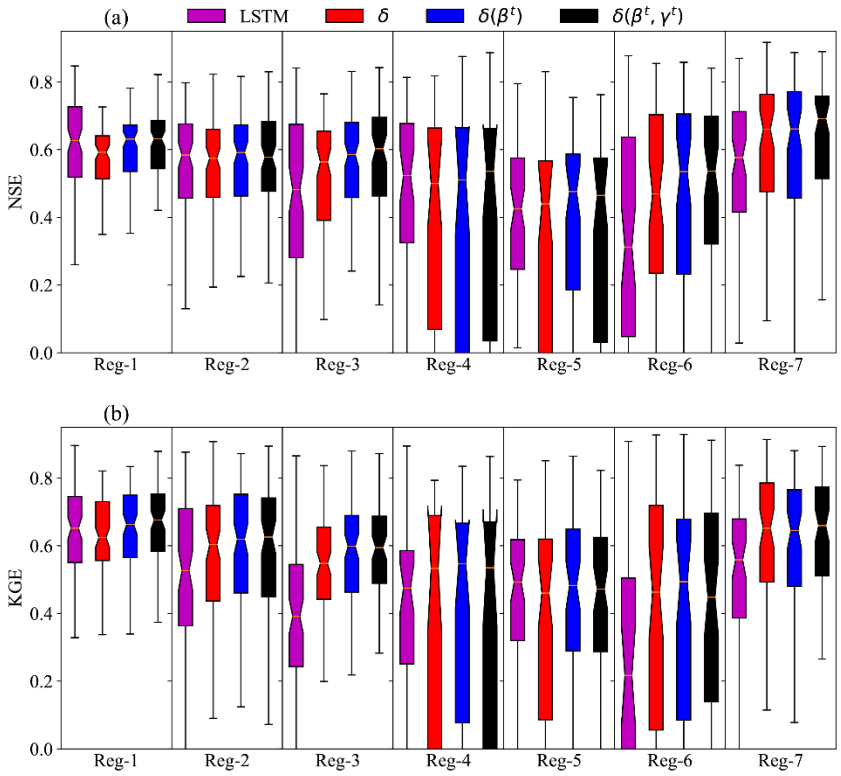
Feng, D., Lawson, K., & Shen, C. (2021). Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data. Geophysical Research Letters, 48(14), e2021GL092999.

Beck, H. E., Pan, M., Lin, P., Seibert, J., van Dijk, A. I., & Wood, E. F. (2020). Global fully distributed parameter regionalization based on observed streamflow from 4,229 headwater catchments. Journal of Geophysical Research: Atmospheres, 125(17), e2019JD031485.

Line 308: Breaking down Figure 6 by region would add a lot more value to the results. This would be super valuable for understanding some of the regional trends in model performance in general, particularly in Regions 4 and 5.

Thanks for this good idea and we added below figures showing the NSE/KGE metric in all the seven PUR regions as well as the related discussions in the revised manuscript.

"With the exception of Regions 4 and 5, the δ models have advantages over LSTM in nearly all other PUR regions, suggesting that the benefits of physical structure for extrapolation are robust in most situations (Figure A2). Region 5 is the Southern Great Plains, with frequent flash floods and karst geology, and both types of models performed equally poorly. δ(βt,γt) showed significant performance advantages in Regions 3, 6 and 7. It is unclear why larger differences exist in these regions rather than others. We surmise that these regions feature large diversity in the landscape (as opposed to Regions 2, 4, and 5, which are more homogeneous forest or prairie on the Great Plains), which when missing from the training data could cause a data-driven model like LSTM to incur large errors. Meanwhile, all the models achieve their best PUR results in Region 1 (Northeast) and Region 7 (Northwest) with NSE/KGE medians larger than or close to 0.6 (Figure 7), which are consistent with our previous PUR study using LSTM (Feng et al., 2021). We also observe that both LSTM and HBV models have difficulty with accurately characterizing hydrologic processes in arid basins as shown by Regions 4 and 5 in the middle CONUS." {line 360}



REFERENCES:

Le, M.-H., Kim, H., Adam, S., Do, H. X., Beling, P., and Lakshmi, V.: Streamflow Estimation in

Ungauged Regions using Machine Learning: Quantifying Uncertainties in Geographic Extrapolation, Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/hess-2022-320, in review, 2022.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. Water Resources Research, 55(12), 11344–11354. https://doi.org/10/gg4ck8

Kratzert et al., 2021. A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling. https://doi.org/10.5194/hess-25-2685-2021

CC1: 'Comment on hess-2022-245', John Ding, 12 Aug 2022

An autoregressive process of the streamflow as a candidate model

The paper presents results from a model comparison of an LSTM vs. HBV and its two surface-runoff-storage variants called the delta models and having one or two time-dependent "dynamic parameters." Their Figure 3 for PUB (B for basins in Prediction for Ungauged Basins) and, especially, Figure 6 for PUR (R for regions) call into question a prevailing claim about the superiority of the LSTM in hydrology, Mai et al. (2022) being a latest.

To cover the spectrum/range of hydrologic models, the authors may want to include one from time series models, such as autogressive processes of (only) the streamflow.

I suggest the authors consider a simplest AR(2) model, a second-order autoregressive process of the form, e.g., Mizukami et al. (2021, SC1 by Ding therein):

*Qsim[t+1]=2.0*Qobs[t]-Qobs[t-1].*

This has been put forward as an alternate reference or baseline model to the observed mean flow one in the popular though rudimentary NSE (Nash-Sutcliffe efficiency) criterion. Azmi et al. (2021, SC1 by Ding & AC1 therein) showed this a good performance model.

I look forward to seeing an expansion of Figures 3 and 6 in a future study covering the AR(2).

References

Azmi, E., Ehret, U., Weijs, S. V., Ruddell, B. L., and Perdigão, R. A. P.: Technical note: "Bit by bit": a practical and general approach for evaluating model computational complexity vs. model performance, Hydrol. Earth Syst. Sci., 25, 1103–1115, https://doi.org/10.5194/hess-25-1103-2021, 2021.

Mai, J., Shen, H., Tolson, B. A., Gaborit, É., Arsenault, R., Craig, J. R., Fortin, V., Fry, L. M., Gauch, M., Klotz, D., Kratzert, F., O'Brien, N., Princz, D. G., Rasiya Koya, S., Roy, T., Seglenieks, F., Shrestha, N. K., Temgoua, A. G. T., Vionnet, V., and Waddell, J. W.: The Great Lakes Runoff Intercomparison Project Phase 4: the Great Lakes (GRIP-GL), Hydrol. Earth Syst. Sci., 26, 3537–3572, https://doi.org/10.5194/hess-26-3537-2022, 2022.

Mizukami, N., Rakovec, O., Newman, A. J., Clark, M. P., Wood, A. W., Gupta, H. V., and Kumar, R.: On the choice of calibration metrics for "high-flow" estimation using hydrologic models, Hydrol. Earth Syst. Sci., 23, 2601–2614, https://doi.org/10.5194/hess-23-2601-2019, 2019.

Citation: https://doi.org/10.5194/hess-2022-245-CC1

AC7: 'Reply on CC1', Chaopeng Shen, 04 Dec 2022

Thanks a lot for the comments regarding autoregressive models. Since the main topic discussed in this paper is the prediction in ungauged regions which assume there are no observations in the target regions, we don't think autoregressive models with observations at previous time steps as inputs are appropriate for the topic. Moreover, we have already compared the deep learning LSTM models with AR models for streamflow forecasting in our previous studies (please see Table 3 in Feng et al., 2020, and also Fang et al., 2017), which has shown deep learning models can largely outperform AR models for integrating historical observations.

Feng et al., 2020. https://doi.org/10.1029/2019WR026793

Fang et al., 2017. http://onlinelibrary.wiley.com/doi/10.1002/2017GL075619/full

Citation: https://doi.org/10.5194/hess-2022-245-AC7