Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-221-RC2, 2020

Hydrology and Earth System Sciences
Discussions

Interactive comment on "A note on leveraging synergy in multiple meteorological datasets with deep learning for rainfall-runoff modeling" *by* Frederik Kratzert et al.

Anonymous Referee #2

Received and published: 9 July 2020

The main point of the paper was that one can use multiple precipitation products in a single LSTM to improve streamflow model performance. The other analyses are secondary (and also problematic), without memorable take-home messages. One of the two conclusions in the abstract is "A sensitivity analysis showed that the LSTM learned to utilize different precipitation products in different ways in different basins and for simulating different parts of the hydrograph in individual basins", which does not seem to have said much.

I think there is some value in this idea (although incremental) of fusing multiple forcing dataset using DL, but the effectiveness of DL has not been compared to other methods

and is thus out of context. Also, stacking multiple data sources as inputs is a common practice among machine learning practitioners. The procedure itself is not the novelty, although I get it that in hydrology few might have used it.

This might be a matter of personal opinion, so I ask the editor to weigh my opinion as what it is — an opinion, the paper appears too thin and too incremental for me to warrant a HESS contribution. This is based on my understanding of HESS as a premier outlet for hydrological science. When the authors published their first couple of papers, which they cited, it was novel. Now, LSTM seems to be widely used in hydrology and it is no longer novel, so the sole point become the use of multiple forcing datasets. My personal judgement is that this point alone lacks the punch needed for a HESS paper.

Beyond the main opinion, I raise some other major points below. It does occur to me many of the claims were rather casually made in this paper and need further validation. Many details were missing, and I would be worried some results are not stable.

1.The motivation seems problematic in logic. What is such a model used for? (i) Are you using it for climate change impact assessment? You are not going to have three forcing datasets which you can train the model with. None of the datasets will be available for future climate. Climate model outputs are not able to be used in supervised training like this, with daily streamflow as the target. (ii) Are you using it for flood forecasting? It does not seem like this model is optimally wired for forecasting, which in general taps into data assimilation. It is uncertain is significant value of multiple forcing would still exist in a setting with data assimilation. (iii) Are you using it for hydrologic budget analysis? With this setting, we don't even know how much rainfall has been applied in the model from a mass balance point of view. Hence, while the results may look nice, it may not have real-world use cases!

2. I wish not to see hydrology becoming a computer-science competition where an incremental change in an experiment becomes a new paper. There should be either scientific advances or methodological innovations. The results might be publishable

but as it reads now it does not look to be at HESS level.

3. The authors claimed DL is better than traditional hydrologic models at using multiple sources of information. This is not proven. One can run an ensemble of simulations with multiple forcings, as authors said around line 270. Now, I do expect DL models to outperform, because previous papers have shown that. but perhaps the difference between single-forcing and multiple-forcing will be similar to what is shown here. At least no hard evidence was provided.

4. Does the author use only multiple precipitation data or all the forcing variables? If only multiple precipitation were used, the author should clearly state this and use "multiple precipitation" instead of "multiple meteorological/forcing" in the title and main texts. If multiple datasets were used for all forcing variables, why were the analyses only executed on the precipitation in section 3.2 and 3.3? How did you know the effects were not due to other variables? If only precipitation was used, why not use other variables?

5. The benchmark scenarios with the hydrologic models are off topic. The comparison between the LSTM and these hydrologic models have been done in Kratzert et. al, 2019. Therefore, it's quite obvious that the multi-forcing model in this study would further outperform hydrologic models. Given the main topic here is to show the effective synergy of multiple meteorological forcings using deep learning model, the fair benchmark would be the ensemble forcing simulations with hydrologic models.

6. The explanation of the triple collocation method was not clearly provided in section 2.4.2. The meanings of α, β, e were not explained in equation 1, 2 and 3. How was Equation 4 derived from the equation 2 and 3? I was confused here. Did the author first fit a log-transformed linear model (equation 2 and 3) or directly use the covariances of time series to calculate the error variance as shown by equation 4?

7. The analysis of gradients is very unclear. In section 2.4.3 and 3.3, why did the author choose the "integrated gradients" instead of the simple gradients of inputs extracted

from the network? At the time of simulation, only the actual variable value matters, not the whole range of x. And... how did you integrate it for different values in x? It wasn't clear at all. If the author chose a more complex implementation, the comparison with the original one is needed. I believe most readers will have hard time understanding what the "gradients" really means as shown in section 3.3. For example, precipitation at previous hundreds of days (i.e. the length of training instance) can all contribute to the runoff prediction at the present time step T because of the memory characteristic of the LSTM model. Did the gradient at time step T refer to the gradient of the precipitation at the present time step or the sum/average gradients of all the previous hundreds of days? More clarifications should be given to help readers understand the gradient results here. 8. Following the above comments, the gradients of inputs w.r.t. the outputs can be quite unstable for deep learning models. Were the results shown in section 3.3 based on the gradients of ensemble runs or single model? If single model was used here, the author should show the results of multiple ensemble members as well as standard deviations and demonstrate the robustness of their gradients.

9. The analysis in section 3.2 is not convincing: 1) In line 210-212, the author clearly stated that the model performed worse in basins with lower precipitation error, especially for NLDAS forcing. This is counter-intuitive, leading me to question the validity of the triple collocation method employed and what it really says. How can the author verify the validity of this method given this abnormal result? Although they have listed the locations of those abnormal basins, they did not give convincing and detailed explanations for this problem. 2) What are the differences of "σ" and "total variance" in Figure 7 and 8? More explanations are needed here to avoid confusion. 3) Line 218-219, "with higher total precipitation variance (not triple collocation error variance), indicating better performance in wetter catchments." This expression is not rigorous since the arid areas with frequent extreme events can also have large precipitation variance, such as Texas. 4) Line 220- 221, "This is not true for the other two models, where higher total variance is associated with a higher variance in model skill, indicating higher proportion of the variance due to measurement error". This is very hard to understand and needs

further supporting evidence. I can only see that the variance spread is large for basins with high NSE performance for Maurer and NLDAS from Figure 9.

5) The same problem as the point 1) in line 228 and Figure 10, the author can not draw a general conclusion and neglect those abnormal basins in Figure 7 and 10. These basins behave differently than the conclusion, so it might be your conclusion that is wrong!

Other comments:

1) Why were only 447 of the 531 basins used for the benchmark with hydrologic models (section 2.4.1)? I went back to their 2019 paper and they used 571 there, where they benchmarked against other models. Furthermore, as we discussed earlier, there is no longer a point to benchmark against default traditional models. It has been done. I would welcome a benchmark with the ensemble forcing scenario, but that was not included.

2) The paragraph starting at line 43 cited the statements in Behnke et al. (2016) and seemed quite incoherent here. It seems this paragraph should be better moved to the introduction part.

3) Figure 7 and 6 can be combined since they tell similar stories. There have been so many figures in the paper which made the paper look redundant. Really with the actual content available in this paper 4 figures would have been adequate.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-221, 2020.