Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2022-77-RC2, 2022

Comment on hess-2022-77

Anonymous Referee #2

Referee comment on "Using LSTM to monitor stormflow discharge indirectly with electrical conductivity observations" by Yong Chang et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2022-77-RC2, 2022

Yong Chang et al. present a study on estimating hourly discharge in a small 1 km^2 karst catchment from precipitation and EC measurements using a LSTM. They set up three different LSTMs based on EC, precipitation and both signals together. Moreover, they explore the performance of other versions of these models with a reduced amount of provided training data. The topic of the study is an interesting contribution to the field, since the added value of EC measurements with gauge levels is indeed underexplored. Also the question about gauging strategies to build a rating curve is of interest. However, I see a couple of severe issues with the study which are in conflict with the strong claims raised and which require to be resolved before final publication.

Major Points

If I understand correctly, the only true estimate of discharge from EC is done with the M_EC model. Given the claims of the title, abstract and introduction, I would not expect precipitation as further variable. L77ff. again precipitation is not mentioned but the use of EC as a proxy. I think the rest of the paper does not really follow this line.

The models including precipitation input are directly predicting discharge. Hence a simple hydrological model and not a linear regression should be the benchmark for these models. Given the situation that the models including precipitation input perform worst in the 2nd evaluation period and otherwise in training and the 1st evaluation period, this raises concerns about what the LSTM actually learned during training. Apparently the temporal patterns of discharge in 2017 and 2019 are more similar than in 2018. What would happen if the model was trained in a different period? Why do the authors expect that the LSTM got sufficient data, when it obviously fails for the test period 2?

Why do the authors use a mean squared error as objective function (L200) instead of a more specific or several complementary evaluation functions?

Using the NSE for evaluation has the known shortcomings and tendency to high values with seasonal climate (Schaefli and Gupta 2007). Given the monsoon climate in the study region, a NSE >0.5 in the evaluation period should not at all be surprising or convincing. Given the adaptability of a LSTM a NSE near 1 should be expected during training. A NSE<0 refers to predictions worse than the mean value. Hence I would expect that the authors would not show arbitrary y-axes limits but to give clear guidance that the performance is not really impressive. Moreover, I would expect further performance measures like KGE, Spearman rank correlation etc.

If I understood correctly, the LSTM is allowed to receive forecasted EC values. I wonder if this is a fair comparison if P is only given in hindcast. If P and EC measurements could be used as proxy measurements, why should I bother about not using forecasted P too? How did the authors assess the chosen time window? I was also unable to identify the m-parameter defining this window. Moreover, I did not really understand the selection of a 7 h time delay factor (L192) since the LSTM should well be capable to learn this.

The authors rightfully expose discharge as central hydrological variable (L36f). But if I would replace this measurement with a model, why should I still be at least somewhat confident about my water balance to be met? Why should I use precipitation as a further explanatory variable to predict discharge if I then would use discharge and precipitation to estimate further characteristics? This fundamentally opens the gates for spurious correlation ill-posing the matter of measuring discharge in the first place.

Given these questions, I am under the impression that the second part of the analyses with different subsets of training data is actually highly case specific. This does not only relate to the selected arrangement of training period, objective function and evaluation procedure. It also refers to the system under study: 1) The authors already modified the EC data (L128ff.). 2) A Karst system should rather directly relate to fill-and-spill dynamics (McDonnell et al. 2020), which are a perfect learning case for LSTMs rarely met in other hydrological systems. 3) The catchment is very small (1 km^2). Hence, I would be very cautious about the capabilities to perform this kind of analysis and the strong claims interpreted from the results. In the current form, I would not really agree that the findings are sufficiently supported.

Minor Points (only points in addition to the major ones are listed)

Title: I find the title not really in line with the content of the paper.

L21: What complex relationship? What special ML architecture? This is far too fuzzy.

L25: I did not spot any assessment of uncertainties. I guess you refer to the overall model performance evaluation.

L39f: depth? water level!; defined relationship? rating curve! Why omitting the established terminology?

Fig 1: I do not really get anything from the maps a and b. Map c is difficult to interpret.

L106: what is a combination of rectangular weirs? Do you have a rating curve for the weirs or is the discharge merely calculated with an empirical weir function? How is the gauge measured? Which uncertainty would you expect?

L108f: I suspect a Onset U24? Why do you report 15 min resolution if later on hourly data is used?

L124: What is unsaturated fast flow?

Fig 2b: Why are the side panels in reverse order and without annotated marks in the main panel?Why is the linear model used as reference not plotted? Why is (again) a different correlation measure used?

L148f: I guess you refer to discharge events (not rain events)?

L155f: A strong relationship? I would not claim a correlation of -0.51 to be specifically strong. Hence the relationship might be somewhat tangible there and is not found when plotting EC to Q for lower discharge.

Sec 3.1: Why dont you calrify your strategy with the three models M_EC, M_P and M_ECP upfront?

L192: What is really meant with the 7h forward shifting?

L206: Why do you report the NSE equation. Not needed. Better add further evaluation estimates.

L263: The benchmark is the linear regression which is slightly better than a pure mean value…

L265: See major points about the NSE and the expectations for an LSTM. Avoid normative claims. Certainly they do not expose excellent capability…

Fig 3: Caption reports Fig 2 instead of 3.

L276: Again, how do you support the claim? Test period 2 obviously fails and it is not analysed if this is due to the lack of precip data. Actually I do not expect that this is the case if the evaluation without OBGD remains that low.

Fig 5: Why do you show Nash values below -1?

[I have not recorded further minor points after L303 since I expect this to require substantial workover anyways.]

Code and data availability: Come on! We are in 2022! I find it absolutely necessary that we do not have to beg for seeing what is under the hood. HESS data and code policy is rather clear about this. I find it as an obligation for the authors to provide their data and code - especially for a study like yours which is merely applying a Keras LSTM so a very limited data set.

——

McDonnell, J. J., Spence, C., Karran, D. J., Meerveld, H. J. (Ilja) van, and Harman, C. J.: Fill-and-Spill: A Process Description of Runoff Generation at the Scale of the Beholder, Water Resour Res, 57, https://doi.org/10.1029/2020wr027514, 2021.

Schaefli, B. and Gupta, H. V.: Do Nash values have value?, Hydrol Process, 21, 2075–2080, https://doi.org/10.1002/hyp.6825, 2007.