Reply on RC2

Yong Chang et al.

Author 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-AC2, 2022

Thank you very much for the valuable comments.

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

1) 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.

Response: Thank you for this remark. The prediction of discharge by precipitation was just used as a comparison to the prediction result by EC. To avoid confusion, the simulation result of model M_{ECP} (using precipitation and EC to predict discharge) will be deleted in the revised manuscript, since it does not provide any effective information in the paper.

2) 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?

Response: We used the linear regression model was used as a benchmark model for M_{EC} since currently there is yet no hydrological model that can predict the discharge using EC. For the model M_p, which uses prediction to predict discharge, we do not apply any benchmarking because this model is just used as a comparison to M_{EC}. We will revise the sentence in lines 189-191.

The models including precipitation, like M_p and M_{ECP} (will be removed in the revised manuscript), has worse performance in the second test period due to a large error of precipitation data (OBGD). Whereas, the performance of model M_{EC} is not severely influenced by the existence of OBGD (see Fig. 3b) because this mode just uses EC to predict discharge. That is, M_{EC} does not fail to predict discharge in the test period 2. This also indicates the advantage of M_{EC} to predict discharge over M_p in mountainous catchments where precipitation has a strong spatial variability. A sparse rain gage network would bring large precipitation uncertainty and bad discharge predictions by M_p.

Since the LSTM is a pure data-driven model, it may have a weak extrapolation ability. Therefore, when the LSTM was used for the discharge prediction by EC, we should collect EC-discharge data under a variety of rainfall conditions.

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

Response: We used the mean squared error as it is a widely-used objective function in many machine learning works, see (Campolo et al., 1999; Gao et al., 2020; Kratzert et al., 2018). The aim of this paper is to explore the feasibility of predicting discharge with EC using a standard LSTM including a typical objective function, i.e. the MSE. Whether the selection of different objective functions affects the final simulation result is beyond the scope of this paper.

4) 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.

Response: We will revise the y-axes limits in Fig.3. In addition, we will provide the KGE and r values of the calibration and validation periods in the revised manuscript. The mean values of KGE of M_{EC} are 0.86, 0.70 and 0.38 in the calibration and two test periods, respectively. The corresponding mean values of the correlation coefficients of M_{EC} are

0.96, 0.82 and 0.73. The low KGE in the test period 2 is due to the poor performance of M_{EC} on the low flows because the low discharge occupy most time in this period.

5) 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.

Response: We only use the previous and current precipitation to predict the current discharge because of the obvious fact that observed spring discharge is just the catchment response to the previous precipitation. The model performance of M_p would not be improved even the precipitation data after the prediction time were used in the model. Whereas for M_{EC}, because the EC dynamic always lags behind discharge, it is necessary to consider the EC data after the prediction time to forecast discharge. The procedure to determine input length (m) is shown in the appendix.

The 7 hours delay was only used in the simple regression benchmark model to account for delay between discharge and EC, not in the LSTM model.

6) 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.

Response: The model M_{ECP} that uses the precipitation and EC to predict discharge will be deleted in the revised manuscript.

7) 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 km2). 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.

Response: Firstly, we would like to clarify that the aim of this paper is to explore for the very first time the ability to use EC to predict discharge using a standard LSTM. Exploring the impact of using different objective functions to train the LSTM would therefore not be the scope of this paper. The longest data series from March 1 to August 1 in 2019 was

selected as the training period since the LSTM is a pure data-driven model and requires abundant data to get a stable simulation result. For the model evaluation, the performance of M_{EC} basically is not influenced by the precipitation error in test period 2 since this model just uses EC as the model input.

Secondly, the adjustment of EC value in test period 1 is based on the fact that the maximum EC of this spring is always relatively stable in different years according to the previous monitoring and different data loggers were used to monitor EC in 2017 and other two years. To further interpret the possible uncertainty caused by this adjustment, we will add another figure to the revised manuscript that shows the variation of model performance with the different EC adjustment values in test period 1.

Finally, this work in the paper is the first time to apply LSTM model to predict discharge using EC. Although the study catchment is small, the observed spring discharge and EC dynamics are similar to many other karst springs (Olarinoye et al., 2020). Therefore, we think the catchment area should not be a problem to apply our approach. Regarding whether our approach can also be used in other hydrological systems, further work is needed which is our next step.

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.

Response: The title will be revised to 'Using LSTM to monitor stormflow discharge indirectly with EC observations' according to the comment from reviewer 1.

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

Response: We will further revise the sentence.

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

Response: Change the word 'uncertainties' to model performance.

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

Response: Accept. We will change the words.

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

Response: Map a shows the location of study catchment in China. Map b displays the locations of two climatic stations and their observations were used to fill two recording gaps. Map c just shows the catchment area of the karst spring.

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?

Response: The discharge is calculated by the empirical weir function. The water level was measured by a HOBO data Logger U20 with precision of 0.3cm.

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

Response: Yes, the Onset U24 was used for the EC monitoring. The hourly data was used because the resolution of discharge in some periods is one hour.

L124: What is unsaturated fast flow?

Response: change to 'low-EC event water'.

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?

Response: We will add the annotation in the main panel. Figure 2b just displays the overall relationship between observed discharge and EC. The different correlation coefficients in the right panel of Fig.2b correspond to a different relationship between discharge and EC under different recharge events. Figure 2 just displays the observation data without any simulation results of different models.

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

Response: Thanks, we will change the words.

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.

Response: We will further revise the sentence.

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

Response: M_{ECP} will be deleted in the revised manuscript. M_p was just used as the comparison to the prediction result by EC. We will add the description in the revised version.

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

Response: The 7h forward shifting was just used in the simple regression model since this model cannot finely learn the delay between EC and discharge.

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

Response: Delete NSE equation.

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

Response: The bad performance of the benchmark model is due to the weak linear relationship between discharge and EC. However, the model M_{EC} can still get a better performance than the linear regression model

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

Response: Excellent changes to 'good'

Fig 3: Caption reports Fig 2 instead of 3.

Response: Thanks.

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.

Response: The bad performance of M_p in the test period 2 even without OBGD is mainly caused by its poor simulation results in the low flow because the low flow takes up the most time in this period with only two storm events. We will add the explanation in the

revised manuscript.

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

Response: Cap the y-axis to -1.

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

Response: We will carefully revise the following sections according to the reviewer's comment before.

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.

Response: The code and data will be uploaded to a public repository.

Reference:

Campolo, M., Andreussi, P., Soldati, A., 1999. River flood forecasting with a neural network model. Water Resour. Res. 35, 1191–1197.

Gao, S., Huang, Y., Zhang, S., Han, J., Wang, G., Zhang, M., Lin, Q., 2020. Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation. J. Hydrol. 589, 125188. https://doi.org/10.1016/j.jhydrol.2020.125188

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall – runoff modelling using Long Short-Term Memory (LSTM) networks, 6005–6022.

Olarinoye, T., Gleeson, T., Marx, V., Seeger, S., Adinehvand, R., Allocca, V., Andreo, B., Apaéstegui, J., Apolit, C., Arfib, B., Auler, A., Bailly-Comte, V., Barberá, J. A., Batiot-Guilhe, C., Bechtel, T., Binet, S., Bittner, D., Blatnik, M., Bolger, T., Hartmann, A. (2020). Global karst springs hydrograph dataset for research and management of the world's fastest-flowing groundwater. Nature Scientific Data, 7(1).