Review on “River flooding mechanisms and their changes in Europe revealed by explainable machine learning” by Shijie Jian et al.

Review by Frederik Kratzert

Summary
This paper presents a large-sample study to detect different flooding mechanisms across Europe. I have to admit that initially, I was skeptical about this study. But while reading the manuscript for this review, I became quite excited about the presented work.

To my knowledge, it is the first time that such an analysis (detecting flooding mechanisms and analyzing the change over time) is made using a) deep learning models (here LSTMs) and b) methods from the field of explainable AI (here integrated gradients). My views on LSTMs is no secret and I have often said that you can do more than “just fitting streamflow records” with these models, so naturally, I am quite excited to see someone coming up with such an idea.

Additionally, I think this paper is exceptionally well written and at least for me personally, everything seemed pretty clear and reasonable. For example, the authors make a couple of assumptions (like grouping the integrated gradient signal into two groups of a) the last 7 days and b) all other days before that), but their reasoning for all these assumptions is clearly articulated and to me, they make sense.

In general, I think this is a very interesting study that fits into the scope of HESS and I only have a few general comments. Note, I already spoke to the first author during the EGU GA but for the sake of transparency, I will add all points here again. Please also note that, due to the overlap of this research with our own research in the past, in many of the studies that I reference I’m either the first or a co-author. I do not mention these studies here, because I want them to be cited in the manuscript, but I think they help to explain my reasoning.

1. **Training setup**
   You train LSTM models individually for each basin, instead of one model on the combined data of all basins. Again, I’m very biased on this topic but I think there are multiple studies that show that the recommended way for training LSTMs is the latter (on all basins at once, using meteorological timeseries features and static attributes). The regional modeling setup was introduced in 2019 (see Kratzert et al. 2019) and further discussed in Nearing et al. (2021) (see Fig 2). One study that follows the regional training scheme is even cited in the manuscript (Lees et al. 2021).
   The question is, is this important in the context of this study? This is a good question that I asked myself quite a lot over the last few days. On one side, I think it is important to

follow best practices when working with any model. The benefit of the LSTM is that it can learn a very general understanding of the underlying processes if it is trained on a variety of basins. Nobody would probably train a conceptual model in a regional calibration scheme, if she/he is only interested in a particular basin. On the other hand, the authors are not interested in getting the best-possible streamflow performance, but to learn about flooding-mechanisms from the model.
From my experience, I would assume that changing the training setup would not change the results of this study (i.e. the clusters of different flooding mechanisms found here). What might change is the number of basins that are considered in their study (because of the NSE threshold). However, even (or especially?) if the results of this study do not change, I would suggest training an LSTM on the combined data of all basins and to re-run the analysis, to reflect the best-practices of the chosen model. During the EGU, I offered Shijie Jiang help with setting up such a run as I have the code + resources available. I would be happy to help and don't want/expect any co-authorship/acknowledgements for that.

2. **Input variables**
   In this study, you only use precipitation, temperature and day-length (as a proxy for solar radiation) as model inputs. I agree that these are the main-drivers for flooding mechanisms but I think the exciting thing about data-driven models is that if there is anything else, these models are pretty good at finding it. That is, if provided with more input features, the models would find other flooding mechanisms, if they are deducible from the data. When I discussed this point with Shijie Jiang at the EGU, he mentioned that he has/had a hard time to interpret the contribution signal for different features, and I agree, it is much simpler and more intuitive to reason about the meaning of high feature importance of e.g. temperature in the recent days. However, how exciting would it be to find that the model finds a flooding mechanism that is not straightforwardly explainable with the patterns we already know? If you have easy access to more input features, I think it could be interesting to run the models with more input features. If not, I think it is ok not to do it, but then it could be worth adding this to the discussion.

3. **Data split**

   In L 138f, you mention that some hyper parameters were determined considering the model performance and efficiency. To me it is unclear which data was used for this hyperparameter tuning. Usually, the validation split (note, here I am referring to a 3-fold split with a train, validation and test set) is used for these kinds of hyperparameter tunings. However, from the explanation in L146, I can only estimate that this was done with a 2-fold split, thus on the test data?

4. **Data split pt. 2**
   In L146, you say that the timeseries data was randomly split in a “7-to-3 proportion”. I am not 100% sure but I think randomly splitting timeseries data is not optimal, especially

when considering the overlap of different samples because of their input window size. I think much more common is a k-fold cross validation in time (none random).

5. **Data split pt. 3**
   I am curious, if you train/eval the models in a 7-to-3 split random fashion, how could you a) guarantee an equal number of model predictions per timestep and b) guarantee that every time step was e.g. evaluated at least once and c) guarantee that the flood peak was in the validation and not the training period? Because from L 169, it seems like you used all 10 models for all peaks, but some peaks are certainly in the training period, right?

References

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, Hydrol. Earth Syst. Sci., 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2021). What role does hydrological science play in the age of machine learning?. Water Resources Research, 57, e2020WR028091. https://doi.org/10.1029/2020WR028091