Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2021-554-RC2, 2022

Comment on hess-2021-554

Anonymous Referee #2

Referee comment on "Flood forecasting with machine learning models in an operational framework" by Sella Nevo et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-554-RC2, 2022

This paper presents ML models that respectively (i) directly predicts the river stage (rather than predicting discharge and then translating to stage); (ii) predict wet/dry of pixels depending on gauge stage; (iii) and estimate flood inundation depth. Among these, (i) was trained based on historical stream gauge data and near-real-time upstream gauges; (ii) was trained using historical satellite data and coincident stream gauge height data. (iii) was not really a model, per say, but an interpolation procedure.

I think the paper demonstrated strong performance from a completely data-driven model. It highlights the idea of directly simulating stream gauge height, which breaks many barriers. If they didn't do this, they need to simulation discharge and then resolve the highly-variable (in space) relationship between discharge and stage height. Most of time we cannot resolve it. In the authors' case, there is no discharge data to begin with. So directly tackling gaging height is a good and necessary idea (but it also leads to some issues I will discuss below). The paper also demonstrates a very efficient forecasting scheme based on upstream gauge data. The whole paper demonstrated how to stack different models together. The authors also showed a unique flood inundation component that is accurate. The work is very useful for hundreds of millions of people and it takes lots of courage to take on such a responsibility.

While there are many reasons why I like this paper and I encourage the publication of this paper, I also noticed a few major issues. These issues are raised here in the hope to make the manuscript more balanced and comprehensive.
(a) There should be some discussion of the potential scientific limitations (even if caused by practical data availability) of the approach and the conditions under which this approach is applicable. As far as I can see, all the models were posed in a highly case-specific way. The gauge height LSTM model has weights that are shared across multiple gauges but it also needs gauge-specific weights that are tuned to local data with a particular configuration. (how much worse will it get if you don't use those gage-specific weights?) The inundation extent model is tied to the gauge and the particular river bathymetry downstream from that gauge. In other words, it seems these models can only be applied where gauge data is available for training. The trained relationship is not

portable anywhere else (if so, it poses a requirement on the available data records). Don't get me wrong. I think the model is highly useful operationally. In India there are many places where the model is applicable. It just might make sense, if these limitations are true, authors can discuss where and when this model formulation is valid so it is easier for the readers to understand if these algorithms are sound for their purpose. Maybe they can come up with a more uniform model and show its accuracy.

This point also contradicts the authors' claim that the model is highly scalable. You cannot take the model to a new terrain and directly apply it. In addition, the learned relationships may not always stand --- what if you have heavy rainfall in the region between your upstream gauge and the gauge of interest? It seems your model cannot consider such forcings (this may not matter that much for large-scale Indian monsoons, but it could be important elsewhere). This means, while the model is fast to run, it is not scalable in the sense of expanding to new areas ---- you must spend the time and effort to collect the data and train the model in every new area of interest, and that is assuming you are lucky enough to have the data. Hence, it is uncertain how the authors intend to use the model on large areas.

It also exerts some constrain on the eligibility of sites. Because you have to train a site-specific model, you can only use sites with long-enough records to train the model. The model cannot be large, and information from other sites do not help with a particular gauge of interest.

If my understanding is incorrect, I stand corrected and the authors can show a test case where the model is applied to an "ungauged" location.

(b) The training dataset for the models were not clearly described. For the inundation extent model, there should be descriptions of how many events were included as training and test images.

(c) It is not clear if the model accuracy drops as we go further downstream from the gauge. Some exploration here will be useful.

(d) regarding authors' criticism on the hydraulic model --- are we sure you feed it the best parameters and inputs? There is no description about calibration. Back to point (a), in a region without past observations, the hydraulic model may still function but the ML inundation model may not --- which means these models have their own use cases. If I'm wrong please correct.

(e) there seemed to be no description of network configurations such as hyperparameters, hidden size, minibatch (maybe there is not a minibatch), training epochs, etc.

(g) does it make sense to average precipitation for a drainage area > 100,000 km^2?

(h) We have no intuitive understanding of what F metrics mean. Do you mind showing some observed vs simulated maps for different values of the F metric?

(i) the flooding depth model was never tested and we do not know its accuracy. Can you talk about its value in the real world? Also, low-resolution could also give you discontinuity.

(j) can this study be reproduced at all? It seems not much of the study can be reproduced or even compared to in terms of data. All the code and data are either proprietary or unavailable. We were just told they could do this and do that and there is no possible path to trying most of the steps here.

Some minor points:

Line 158. What does "State handoff" mean?

Line 190. Should be "Quasi steady state" to be more exact

Line 196. "Discarded" – see my point above, can you use a more gentle word?

Line 198-199. "when the target gauge exceeds a (pixel-specific) threshold water stage. " A bit confused. A gauge is just at one location, then why do you have a pixel-specific threshold linked to a gage? If it is pixel-specific, then you end up getting a map of different thresholds? Should it be image-specific thresholding?

Line 219. Maybe I'm missing sth, although the thresholding model does not need DEM, it is tied to a particular gauge and the particular terrain/floodplain characteristics. It needs to be trained for each domain of interest using historical inundation extent and gauge height data, so it is not clear to me you can deploy to a new region without effort.

Line 375 what happened to the flood and the effectiveness of the alert? You get us concerned but didn't say any outcome.