Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/hess-2022-53-RC1, 2022

Comment on hess-2022-53

Lennart Schmidt (Referee)

Referee comment on "Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States" by Kieran M. R. Hunt et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2022-53-RC1, 2022

General comments

The manuscript delivers an interesting addition to the current surge of machine-learning in hydrological modelling by extending the application of LSTMs from pure streamflow modeling to actual forecasting. To do so, they ingest the output of physical forecasting systems as input to an LSTM. The main result, that LSTM outperforms the other approaches, is not surprising as the LSTM merely acts as a bias correction algorithm with many more degrees of freedom. Nevertheless, this is still a relevant finding that should be disseminated throughout the hydrological community. The manuscript is well-structured and comprehensible, particularly the introduction and methods parts are very comprehensive yet concise. Intuitive measures of model performance, a solid discussion of the error metrics as used in the study, a comprehensible discussion on the nature and choice of datasets as well as informative plots act as a solid foundation for the reader to follow along w.r.t to the methodological execution and its results. However, towards the end, the discussion and conclusion do miss out to put the work and the results into a broader perspective e.g. by contrasting it against ongoing machine-learning research, its limitations or future directions for hybrid modeling (see detailed comments below).

Scientific/Specific comments

1. Please elaborate on the different types of "Hybrid" models/forecasts that are possible. In ML literature, there are current advances, termed "hybrid models", of including and solving differential equations inside the NN, promising the best of both worlds (high accuracy while keeping interpretability/robustness to out-of-distribution cases). These developments should be listed as future directions of research and the approach of this manuscript should be contrasted against these new development in the introduction.

Rackauckas, Christopher, et al. "Universal differential equations for scientific machine

learning." arXiv preprint arXiv:2001.04385 (2020).

Raissi, Maziar, Alireza Yazdani, and George Em Karniadakis. "Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations." Science 367.6481 (2020): 1026-1030. APA

2. As the title suggests, the LSTM-approach is set up to "boost" the forecasts of IFS. As the LSTM receives Era5/IFS streamflow estimates, I personally would rather view it as a more sophisticated bias correction (more parameters, less constraints) than a separate modeling approach. I believe that the manuscript would benefit if this was put into perspective in the discussion part. Also, the possibility of using a simple statistical/ML model with less training effort, e.g. a simple linear model or RandomForest should at least be mentioned as an alternative.

3. Especially when taking into account that 7/10 catchments are known to be too small for the raw GloFAS to perform well, it is obvious that a statistical bias correction outperforms the raw forecast more the more degrees of freedom it is given. The "unfairness" of the comparison of raw GLOFAS vs. LSTM (world-wide simulation at 0.1° resolution vs. local model) should be highlighted in the introduction and discussion sections.

4. It would be beneficial to include and elaborate a bit further the motivation behind the two bias correction algorithms in the introduction to 4.2., i.e. quantile (remap values to reduce systematic bias) and spatial (inherent spatial bias in GLOFIS-ERA5(?)). Also, the motivation for the two final steps should be explained and justified in greater detail. What do you mean by different climatologies, why split 3/4 vs. 1/4? Why do you shift the forecasts that have ben quantile mapped-once more? Also the fact that the bias correction has been newly developed (mentioned in conclusions) should be placed in the respective chapter in the methods part.

5. You argue that you used reanalysis data during train+test to make the results reproducible for potential users. But are the operational forecasts using IFS still reproducible? If not, it would be beneficial to provide the respective data on zenodo or a similar platform.

6. Generally, the manuscript misses to give detailled information on the training process. I would advise to include loss curve(s) (loss vs. epoch) of test and train. This is the common way to present information to see whether training was successful. Also, train and test error metrics should be provided to give an intuition whether under- or overfitting might have happened. The same applies to the bias correction, here the reader is not provided with any information on the optimization procedure or performance, even though there is a risk of overfitting. Similarly, information on the loss function, training hyperparameters (dropout, decay, learning rate, recurrent activation etc.) should be listed in appendix or refer to repository. To this end, the training process could have been performed more thoroughly: Hyperparameter-Tuning usually involves searching over

model/training parameters as well as model configurations (n hidden layers, nodes etc.), not only epochs. The latter is usually less relevant. Ideally, hyperparameter tuning would be executed in a cross-validation set-up. To that respect, please elaborate what "tuned using sensitivity tests" refers to (l.240).

7. The conclusion is (too) detailled on the technical side (n of skilful vs. non-skilful results, KGE vs. NSE) but misses to provide the most relevant point in 2-3 comprehensive sentences: Where exactly does LSTM perform best/worst (seasonal, diurnal, altitude etc.), Where are its limitations? Both the discussion and conclusion miss to put the results into a broader perspective: How could the LSTM be improved other than switching to Convolutional LSTM-layers? What are future research directions? Maybe some points to consider here:

- Limitations of LSTMs: Some limitations are fairly well-acknowledged by now (parallelization, long-range dependencies) so that LSTMs are not the go-to model for sequential data anymore. Thus, outlook on new developments like GRUs and, most importantly attention-based models (e.g. transformers) should be included
- Limitations of ML generally: Generalisation when provided with out-of-distribution data, e.g. due to system changes (climate change), lack of interpretability
- Greater picture:
  - Comparison to other hybrid modelling approaches in ML (see above)
  - How far away are we from entirely ML-based forecasts, given considerable advances in ML-based climatological forecasts? Should science still focus on improving relatively coarse physical model like ERA5 or rather explore ML-based bias correction at a large spatial scale?
- Reflect on the risk of overfitting in bias correction + lstms, including the fact that, ideally, one would have to estimate the bias of IFS w.r.t to ERA5.

Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Technical Corrections

- The abstract could be shortened
- l. 85, add bracket
- l. 110: "interesting challenge" is a subjective statement, rephrasing is advised
- l. 252 wether A extending
- l.262 remove note
- l. 281: Is it 6 or 7 catchments that are skilful? 5.1.1. lists 6/10 that are skilful, here it is "still" 7/10?
- l. 312 remove (CHECK)
- caption fig. 9: "black" is actually grey
- Please elaborate what the control member in GLOFAS/IFS (ll. 129;467) and the "ensemble member" of the LSTM are (fig. 5)

- The detailled description of and motivation behind the choice of training, testing and operational timespans could be placed elsewhere than in 4.2.1. Possibly best at beginning of chapter 4. Please also make it clear that LSTM and bias correction use the same time spans.

Potential changes for readability:

- 4.3. Include input features as table, not in text
- Present accuracies (aka skilfulness) of the three models as a table