# Review of ESSD-2022-52: WaterBench: A Large-scale Benchmark Dataset for Data-Driven Streamflow Forecasting

In their manuscript, the authors present a hydrological benchmark product particularly for ML-approaches. They've collected hydrological time-series data for precipitation, streamflow, evapotranspiration as well as static data for soil types, slopes, etc. mainly for the state of Iowa. This data is area-averaged over several small basins and arranged in a way, that each of their files contains all data for a single sub-basin. They also include a file that describes the relationships between the different sub-basins.

Using this data, they compute and evaluate some 5-day-forecasts by applying several ML-approaches as well as Ridge regression.

Overall, I found the idea very interesting and such a well-curated and put-together dataset would be an important contribution to both the hydrological and ML-community; there would also be a lot of potential for enhancing such a product with more variables (e.g., temperature, wind, etc.) and also include other states beyond Iowa.

However, after going through the manuscript and looking at the data, I have several points of criticism. Thus, although I really liked the general idea and also acknowledge the authors motivation to make all data and codes publicly available, I think that the paper, as well as the dataset in its current form, needs some substantial revision. Thus, I suggest to reject the manuscript but also encourage the authors to re-submit a revised version.

## Major comments

- My main concern is that the focus of the paper is not clear. If the authors want to present a state-of-the-art "meteorological forcing" product for testing ML-approaches (which would be very interesting), then I do not understand the choice of datasets (see below) and the structure of the dataset needs some revision (probably as one large NetCDF?). It would also be desirable to provide some alternative data (e.g., different soil maps, different precipitation data, etc.) in order to run, e.g., ensemble experiments. However, if the authors rather focus on the development of ML-based streamflow reference forecasts, against which other ML-approaches can be tested, then the methods-section as well as the uncertainty-analysis needs to be substantially enhanced. And, even if the authors claim several times that flood forecasting is an important task (which requires very detailed and precise evaluations), the results-section is restricted to a very high-level comparison of median NSE and KGE values across all the 125 stations.

- Overall, the whole presentation of the ML-predictions is lacking a lot of details. Thus, even if the authors state that their predictions are just examples what one could do with their dataset, there is basically no methods-section and no discussion, why the chosen ML models might be appropriate for flood forecasting. Instead, this whole experiment seems to be another example of a more or less loose conglomerate of ML-approaches. Due to the extreme frequency where new and even more user-friendly ML-libraries are released practically every day, there are more and more papers where people simply use these methods because they can! Here, the authors simply state that they're using Ridge

regression, LSTM, GRU and S2S. It remains unclear why (or why not) these approaches are particular suited for this application and also the reasons for the differences are left completely open.

- Dataset / manuscript currently does not comply with ESSD-regulations:
  - Your dataset needs a DOI (https://www.earth-system-science-data.net/policies/data_policy.html)...
  - ...which you can obtain by uploading to a long-term repository (https://www.earth-system-science-data.net/policies/repository_criteria.html
  - This DOI should also be added to your abstract (https://www.earth-system-science-data.net/submission.html)
  - Furthermore, while it is highly acknowledged that the authors have made all their code and data openly available, I found especially the structure of the data very hard to understand!

## Minor comments

- The authors could have chosen more appropriate datasets for the different water-cycle variables (or even provide an ensemble). Especially as the authors apply area-aggregated precipitation and evapotranspiration, there is a huge range of products available. Instead, the authors mix hourly precipitation with monthly evapotranspiration as well as high- with low-resolution data and, hence, provide a quite inconsistent product with a lot of room for improvement.

- The authors cite that "deep neural networks increased scientists' ability in modelling both linear and non-linear problems without time-intensive data engineering processes by domain experts". At another section, they also claim that the application of a particular dataset does not require domain knowledge from meteorologists (line 85). To be honest, I do not think that this is actually a desirable development. I even think that some proof-reading from a "domain native" could have substantially improved the paper as the authors have, at several places, chosen some quite strange and uncommon wording (see particularly the description of the different variables section 2; some examples are given below).

- It is also not clear how the authors defined their benchmark setup. I assume that they made 5-day-predictions on every day during their test year. From these predictions, they combined all forecasts that refer to a specific lead time (e.g., 1 hour, 6 hours, etc.) into a single sample, over which they then compute the NSEs and KGEs. In their figures and tables, they, finally, present the median NSE and KGE values over all 125 basins. Is that, more or less, correct?

- Lines 33 - 34: I am not sure if "flood forecasting" is really a synonym for "streamflow prediction / runoff forecasting". Furthermore, "runoff" usually describes water on an area that does not infiltrate or evaporate and, hence, discharges from that area; streamflow is the actual discharge that you measure in a river or channel. Thus, while these terms are very related, they do not mean the same thing.

- Line 69: I found the wording here quite strange. You usually do not forecast measurements, but some phenomenon like precipitation or runoff.

- Line 120: "The WaterBench is not selected based on human activities, which is a reaction to the real situation in Iowa" --> What do you want to say here?

- Lines 139 - 144: This sounds not very "hydrological". You should rather say that red dots are located at outlets of larger basins, which are divided into several smaller upstream sub-basins. And the outflow from each sub-basin is measured at the green gauging stations. See, e.g., http://proceedings.esri.com/library/userconf/proc01/professional/papers/pap1008/p1008.html for a good explanation of the terminology. For better visibility, it might make more sense to separate the larger basins from their sub-basins, e.g., by using thick and thin lines in Figure 1.

- Line 148: What are "statistics of the metadata"?

- Line 149 - 150: "Metadata" usually refers to "data that provides information about data". You mean simply "static data". Please rephrase: For each basin, we provide static data (area, slope, travel time, ...) as well as time-series for streamflow, precipitation, and ET.

- Line 195: Your particular soil dataset has 12 soil-types. And I am pretty sure, that there are newer and much higher resolved maps available (see, e.g., https://www.cen.uni-hamburg.de/icdc/data/land/soilmap.html). Such data would fit better to your high-resolved precipitation data (even if you're only looking at basin averages).

- Line 206: Please re-phrase: For each station, streamflow data was aggregated to hourly values.

- Line 213: This sounds, once more, very "un-hydrological" and I strongly assume that you don't have to explain what "precipitation" is.

- Lines 206 - 207: Since there were... -> Please check your grammar... This sentence does not make sense.

- Lines 218 - 220: This sounds, again, overly complex... For "precipitation on the watershed", you usually say "basin-averaged precipitation".

- Lines 222 - 223: Describing evapotranspiration as a major loss of precipitation sounds quite uncommon. Maybe say "precipitated water".

- Line 223: "no high-resolution ET dataset" → What about ERA5 (Land)? GLEAM? MERRA? MSWX?

- Line 224: What do you mean with "empirical dataset"?

- Line 252: Please rephrase: where Y_i is the observation at time i, Y^hat_i is the model result at time i, ...; and please add that sigma and mu refer to the forecasts while your Y refer to the observations! Furthermore, you usually use the same parameter "family" for the mean and standard deviation. So, either use Greek letters (as, e.g., in

https://hess.copernicus.org/preprints/hess-2019-327/hess-2019-327.pdf), or stick to Latin letters (as in your Equation 1).

- Line 258: "Thus, a median value...": While this is certainly true, it also makes a lot of sense to analyse the distribution of your performance metrics across your 125 basins in order to get an idea where and why your model performs better/worse. Figure 4 is only a starting point for such an analysis.

- Line 260: Why are the 120hr ahead predictions the most important values?

- Line 297: I would consider 5-day-forecasts as "medium-range

- Figure 2: Presenting CDFs for static parameters is quite unusual. It would be more helpful if you show Histograms here.

- Table 3: You would rather say "Summary statistics for precipitation and streamflow". And, for better readability, please add the period during which these numbers were calculated.