Earth Syst. Sci. Data Discuss., author comment AC3
https://doi.org/10.5194/essd-2022-52-AC3, 2022


# Reply on RC2

Ibrahim Demir et al.

---

Author comment on "WaterBench-Iowa: a large-scale benchmark dataset for data-driven streamflow forecasting" by Ibrahim Demir et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-52-AC3, 2022

---

**Reviewer 2 Comments and Our Reply:**

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.

Answer: Other variables such as temperature, wind speed, wind direction, snowfall, and snow cover were applied in our preliminary research. However, those variables did not improve the performance of the model performance of the 120hr streamflow forecast. In contrast, a powerful data-driven model may easily overfit to features that are not statistically associated with short- or medium-range floods. In addition, one of the goals of our dataset is to be used directly by computer scientists or data scientists who are not familiar with the hydrology domain. Thus, we limited our features so that scientists in other domains could also use the dataset. Due to our regional focus on Iowa flood studies in data collection, we have renamed the title of the manuscript to reflect our spatial focus.

**Major comments:**

1. 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.

**Answer:**

We used the domain knowledge of hydrology to preprocess the hydrological data in various formats (e.g., grib2 and NetCDF), and summarized them in the general data format CSV which is a widely used format by data scientists. We have tried a variety of rainfall products in our preliminary experiments, and this set of data we provided has been proved to be effective in Iowa studies. Our goal is to enable more computer scientists and data scientists who lack the hydrology knowledge to develop data-driven models on our dataset. And this set of ready-to-use data will be helpful for them.

In today's data science fields, there are many competition platforms providing a dataset and evaluation metrics for users to build the model and break the record. And the leaderboard is based on one or more overall scores as the main evaluation metrics. Here are two samples from two famous platforms.

- Study of the impact of air quality on death rates. https://www.kaggle.com/competitions/predict-impact-of-air-quality-on-death-rates/leaderboard

This is a competition launched by the European Centre for Medium Range Weather Forecasts (ECMWF), 2017. This dataset consists of only several CSV files, which simply provide a region ID, mean O3, PM10, PM2.5, NO2, temperature, and mortality rate in a table. The leaderboard is based on the RMSE score of the predicted mortality rate.

- Study of the prediction of monthly air quality in Beijing.

https://paperswithcode.com/sota/multivariate-time-series-imputation-on

This is a dataset that provides 7 features (wind speed, wind direction, rainfall, air temperature, dew point temperature, and air pressure) as input to predict the air quality. The score of the leadboard is MAE of daily PM2.5 prediction among 12 sites.

NSE and KGE are two of the most widely used performance metrics in physical and data-driven hydrological forecast studies. Thus, we are providing the median NSE and KGE as the main metrics in this benchmark.

2. 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.

**Answer:** We have added a new section of the model details and discuss why these approaches are particularly suited for this application. The LSTM, GRU, and Sequence-to-sequence models on the runoff prediction and streamflow forecast have been widely studied in recent years.

3. 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!

**Answer:** For 3.1, 3.2, 3.3, we will submit the dataset to a long-term repo with a DOI and update the manuscript. The data structure is in CSV files, which is machine learning ready-to-use structure and can be read as a data frame directly. Our paper mainly described the raw data. For the ready-to-use datasets, we suggest reading through the function ***read_file*** in our sample codes.

**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.

**Answer:** We are providing the same spatial resolution as CAMELS (Newman, 2015), and our benchmark mainly focuses on temporal predictions. There is no real-time hourly evapotranspiration observations, so the mixing and efficient use of different resolutions is also a challenge in modeling and could be an advantage for deep learning. We are trying to provide best available datasets for all physical parameters that are available for supporting real-time prediction. Our goal in creating this task is to implement real-time streamflow forecasts that take advantage of the efficient performance of machine learning. Therefore, we only consider the data that can be obtained during real-time forecasting, including monthly ET that can represent the seasonal changes and so on. Many water cycle variables can be obtained from reanalysis data but not real-time. The original data from multiple resources were inconsistent, but this is exactly our work: our aggregated dataset provides a USGS basin station-level hourly data.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., ... & Duan, Q. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209-223.

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).

**Answer:** We see this as a new kind of collaboration. The raw data is preprocessed by domain experts for some specific tasks, and the tasks are then converted into solving an optimization problem (i.e., language translation models can be applied to the time-series forecast modeling in earth science studies because they are the same sequence step-by-step forecast task [Figure 2d, Reichstein et al., 2019]). It is not an issue if the dataset is correctly processed by domain experts before handed over to data scientists or computer scientists.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N. (2019). Deep learning and process understanding for data-driven Earth system science. Nature, 566(7743), 195-204.

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?

**Answer:** Thank you for the comment. We described the benchmark setup in detail in the revised manuscript. Most current studies (i.e., Kratzert et al., 2019) are using median NSE/KGE over hundreds of basins as the main metric. In addition, as is also mentioned in our answer to your Major comment #1, a simple metric for a task is important in today's benchmarking studies. In this revised manuscript, we also included CDF as you mentioned.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrology and Earth System Sciences, 23(12), 5089-5110. https://hess.copernicus.org/articles/23/5089/2019/)

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.

**Answer:** Thank you for the comment. We have modified the sentence “Streamflow prediction and runoff modeling are modeling efforts where the water from the land or channel over time is being modeled”

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

**Answer:** We have updated to the ”streamflow rate”.

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?

**Answer:** We have now removed this sentence. Some datasets (i.e., CAMELS) only contain the watersheds unaffected by human activity, which is simpler to forecast. However, flood impacts are also important in urban or suburban areas.

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.

**Answer:** Thank you for your suggestions. We updated the sentences and figures.

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

**Answer:** It is updated with “statistics of the data”, including the basic statistical values such as min, max, average, and median value.

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.

**Answer:** Thank you for the suggestion. We have rephrased them.

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).

**Answer:** Your information is very valuable. The soil map you mentioned is more detailed and in a higher resolution. We will verify and test this version of the soil dataset and include it in the future version.

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

**Answer:** We have rephrased this sentence as you advised.

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

**Answer:** We have removed this sentence.

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

**Answer:** We have rephased this sentence. “The original data contains a few missing values due to station system failures or internet outages.”

Lines 218 - 220: This sounds, again, overly complex... For "precipitation on the

watershed", you usually say "basin-averaged precipitation".

**Answer:** We have revised this sentence as you advised. “In the dataset, we provide the hourly basin-averaged precipitation data for each station.”

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

**Answer:** We have revised this sentence as you advised.

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

**Answer:** We have revised this to “no high-resoluation real-time ET dataset”. Our final goal of the expected data-driven or deep learning model is to make real-time predictions. Thus, we did not include the reanalysis model output data that we cannot obtain in real-time. For example, ERA5-Land has a delay of 3 months, and GLEAM is only available till the end of 2021 (around 6 months delay).

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

**Answer:** For the evapotranspiration, the IFC has been using a simple climatology based on 12 years of North American Land Data Assimilation System. This approach captures the seasonal effects but fails to account for year-to-year and day-to-day variability.

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).

**Answer:** We will revise this sentence as you advised.

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.

**Answer:** We provided the standard deviation in the updated Table 4. We also included the CDF now. This is a benchmark paper to provide basic results, and we encourage the future researchers to beat our benchmark results with higher median NSE or KGE.

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

**Answer:** It is easy to predict the streamflow for the next 1-6 hours since it would not change too much from the streamflow rate at hour 0. The model performance always decreases with the prediction time. Thus, the model efficiency at the 120hr ahead prediction is the most important.

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

**Answer:** We have updated it to medium-range in this manuscript as you advised.

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

**Answer:** We have updated the CDFs in Figure 2.

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.

**Answer:** We have revised the table as you advised. It is now “Summary statistics for precipitation and streamflow among 125 catchments from water year 2012 to 2018.”