General comments

This paper proposes a comprehensive reference dataset for streamflow prediction based on use in data-driven and machine learning studies, providing reference performance for more advanced deep learning architectures on datasets for comparative analysis. There is certainly a lack of unified reference data in earth science research, and these studies are very useful in being able to outline a general algorithm for defining them. A detailed description of the results is given in the study, but a more detailed description on the data, but also on the climatology of the study area should probably be reported to provide more details for researchers who intend to use WaterBench for deep learning research in hydrology.

Specific comments

Line 17: *"To some extent, WaterBench makes up for the lack of unified benchmarks in earth science research. We highly encourage researchers to use the WaterBench for deep learning research in hydrology."*
Suggestion: I would replace this sentence with a summary of the results; it will be the good performance that will encourage.

Line 84: "*This dataset solves the difficulty of data acquisition and does not require domain knowledge from meteorology."*
Question: What the authors mean by this sentence: "*… does not require knowledge of the domain from meteorology?*"

Line 89: *"Scientific advancement, intrinsically, is supposed to be cumulative, and in order to have better generalized deep learning-based flood forecasting models, scientists need to build on top of what their fellow researchers have done."*
Comment: Certainly, scientists need to build on what their fellow researchers have done, but they also have to improve on what has been done.

Line 90: *"We believe that this could only be done by using the same set of testing mechanisms …"*
Comment: This sentence is true after being certain that everything has been done correctly.

Line 102: *"There remains a need for a dataset that is more convenient to use in deep learning research given that most of the deep learning researchers are not domain experts."*
Question: What the authors mean by this sentence?

Line 110: *"This study proposes a flood forecasting dataset that follows FAIR data principles …"*
Suggestion: I would write some references and details about FAIR.

Line 115: *"… it could be used in other studies with similar goals, such as physically based modeling with physical equations."*
Suggestion: I think "*with physical equations"* is not needed.

Line 165: *"From the tables above, it is shown that our dataset is limited to a certain range of precipitation since it contains the catchments only in Iowa."*
Question: What the authors mean by this sentence?

Line 167: *"Geologically, all the catchments are located in two HUC watersheds, the Upper Mississippi and Missouri, and the study results may not be applicable to other regions in the U.S."*
Question: Do the authors think that this study is only about the analyzed area?

Line 168: *"WaterBench is also subject to a relatively high missing data rate for streamflow since the reliable hourly dataset is limited in USGS for some of the watersheds in Iowa."*
Question: How have you thought about solving this limitation? How much data is missing?

Line 174: *"Since the total precipitation amount is the product of precipitation intensity and area, larger watersheds typically have higher streamflow rates."*
Comment: I do not think this comment is needed, because river flow discharge depend on a number of conditions.

Line 179: *"2.2.2 Time of Concentration."*
Question: Isn't the time of concentration calculated for each grid point of the considered domain? Is the velocity not estimated as a function of slope?

Line 188: *"A steep slope may cause a higher velocity and lower infiltration rate, which normally causes a larger streamflow rate at a precipitation event"*
Suggestion: Therefore, it would be very important to have the velocity data for the different considered grid points.

Line 206: *"Since there were a few missing values in the original data caused by station system breakdown or internet outages."*
Question: How is missing data handled?

Line 221: *"2.2.7 Evapotranspiration (ET)."*
Question: Do the authors believe that this comprehensive reference dataset for streamflow prediction is intended for climate or meteorological studies?
For climate studies the monthly ET parameter may be useful; for meteorological-hydro studies perhaps, it would be more functional to use soil moisture data (also useful for climate studies) and at a higher temporal resolution.

Line 145: *"Since general statistics such as mean squared error (MSE) and root mean squared error (RMSE) are not dimensionless…"*
Suggestion: However, it would be interesting to see some results.

Line 255: "*This means that there will be 120 different NSE and KGE values for different hours at each watershed*."
Question: What the authors mean by this sentence?

Line 256: "*It should be noted that since the watersheds here are not filtered, it is possible for some watersheds to be greatly affected by human activities, including mitigation, construction, irrigation, urban drainage, etc. activities in watersheds.*"
Question: Doesn't this sentence conflict with what is stated in line 120?

Line 275: "Table 4. The median NSE and KGE among 125 watersheds at the prediction hour 1, 6, 12, 24, 48, 72, 96, and 120*.*"
Question: what do the second and third columns indicate?

Line 287: "Table 5. The median NSE and KGE among 125 watersheds at the prediction hour 1, 6, 12, 24, 48, 72, 96, and 120*.*"
Question: what do the second and third columns indicate?

Line 297: *"The second limitation is the scale effect."*
Comment: Unfortunately, it is not clear from the work what scales are being considered. Would the scale effects be the size of the drainage basin?

Line 300: "The scale effect observed in our benchmark indicates the prediction on small watersheds is still a challenge*.*"
Question: Given the Classification of Watersheds by Size from the values in Table 2 it is evident the small to medium sized watersheds in the considered domain are a large number. I do not understand the statement made in line 300.

Line 315: *"Although the data is limited to the Midwest, we believe that any studies on this dataset could provide insights for other streamflow forecasting and rainfall-runoff modeling studies at other watersheds."*
Comment: This sentence written in this way is contrary to what is stated in line 168.