Review of “HERA-WB: a multidecadal, sub-daily water balance dataset for Europe (1951–2020)” by Tilloy et. al.

This manuscript presents a catchment-level dataset of 17 blue and green hydrological variables for 2,214 European catchments, covering 1951-2020 at high spatio-temporal resolution. The data are simulated with the open-source LISFLOOD hydrological model, forced by bias-corrected ERA5-Land climate reanalysis. The dataset is a valuable resource for the community. It can support the development and benchmarking of machine-learning hydrological models, and it enables studies of blue and green water resources across Europe.

I have a few comments that, in my view, would increase the dataset’s usefulness.

A large number of catchments show weak, and in some cases negative, Spearman rank correlations for evapotranspiration (Top20, Bot20) and surface soil moisture (daily, Top20, Bot20). Even where these correlations are statistically significant, their low values suggest the model may miss important local hydrological dynamics.

I see two ways to make the evaluation of these variables more robust. Use multiple reference datasets. Evaluating against a single product implicitly treats it as the “true” target, even though every product carries its own uncertainty. Agreement or disagreement with one dataset therefore says little about model accuracy. I recommend showing that the model output falls within the observational uncertainty envelope of several independent datasets. For Surface soil moisture author’s can compare against a combination of SMAP, SMOS, and AMSR-E/AMSR2 or even ML based products. Similary for Evapotranspiration extend the comparison to products such as MOD16 and HOLAPS, or other available flux products. These satellite products do not cover the full 1951-2020 period, but they can still be used over their periods of overlap with the dataset. Demonstrating that the model is physically plausible and performs within the limits of current observational uncertainty across three to four datasets would be a convincing validation, even if applied only to soil moisture and evapotranspiration. Second, Spearman correlation measures only the consistency of ranking and says nothing about the magnitude of the simulated values. A model can have a high rank correlation and still be strongly biased, or vice versa. I suggest complementing the rank correlations with magnitude-based metrics.

The dataset provides precipitation (P), evapotranspiration, and runoff (R), but each is evaluated independently against a single reference product. Independent evaluation is useful, but it does not show whether the dataset partitions precipitation well enough between blue and green water. Compensating errors can hide this: an overestimate of evapotranspiration and an underestimate of runoff can each look acceptable in isolation while the partitioning is wrong.

I recommend that the authors calculate a blue-green water share index and compare it with an observation-based estimate across Europe. One example is the Blue-Green Water Share (Heselschwerdt et al., 2026), where BGWS = (R - E) / P × 100, where R is runoff, E is transpiration, and P is precipitation. The authors are free to choose a different index if they consider it more suitable. Such an index turns three separate checks into one integrated

test of how the dataset balances runoff against evapotranspiration relative to precipitation. I expect it would make the dataset a stronger target for developing and benchmarking machine-learning hydrological models, and a more reliable basis for analysing changes in freshwater availability over time.

Minor Comments

1. Line 100: Please clarify what is meant by 'focus on high flows'. Does this mean that calibration with KGE is tuned toward high flows, and if so, how does this affect low-flow performance?
2. Line 124: Please clarify how water withdrawal is allocated between sources. Is there a priority order, for example surface water first and groundwater only when surface water is insufficient, or is the demand split between surface water and groundwater according to a fixed fraction?
3. Line 266: What is the minimum record length, and the maximum fraction of missing days allowed within that period?
4. Line 275: I recommend adding a catchment distribution map showing the spatial spread of the 792 evaluated catchments. Also, a station representativeness map displaying the locations of the observation stations within these 792 catchments, using a color scale to indicate the ratio of the station's upstream area to the total catchment upstream area. The catchment with strict matching currently shown in figure 3 are very hard to see. These visualizations will complement the text and give readers greater confidence in the geographic coverage and physical representativeness of the evaluated data.
5. The graphic quality of Figures 1, 2 and 3 can be improved. The annotations are difficult to read, even after zooming in. I recommend increasing the annotation font size. Also, the catchments with strict matching for Q, shown in red, are hard to see.