Review of HESS-2021-279 - How well are we able to close the water budget at the global scale?

In their study, the authors present a comprehensive analysis of the water budget closure over a large number of river basins using a wide range of state-of-the-art global hydrometeorological data. While I have several major and minor comments, I found the manuscript very well written and clearly structured. As the study therefore presents a very comprehensive and detailed overview of current global hydrometeorological datasets and, hence, is a good starting point for future research e.g. on a more regional scale, I suggest a publication after major revisions (see especially my major comments 1 and 2).

Major comments:

- Using the GURN-dataset as the only source of information for runoff is a bit disappointing. It is certainly correct that LSMs without any routing scheme might provide runoff estimates that are not directly comparable to observed runoff on daily or shorter time-scales. But as the authors are analysing monthly and longer-term averages (and even apply a temporal filter), the inconsistencies between observed and modelled runoff should substantially decrease. Moreover, the reasons why we need such comprehensive evaluations and why we see such large discrepancies in the water budgets are that we do not have global reliable information about catchment-scale precipitation and evapotranspiration. But for gauged basins, runoff is usually the most accurately observed variable in the water balance equation. The authors claim that "it is useful to know beforehand which datasets are more reliable to close the water budget in the region under study" (page 11, line 250). And we do have this information for a large number of basins that are analysed in this study. While GURN is certainly a valuable dataset, it is again one step away from the "real" observed quantity, which we, in the end, try to reproduce with all our different global models and remote sensing products.

  Another issue with the GURN dataset is that it is only valid for rivers without extensive human interactions (reservoir management, substantial extraction for irrigation, etc.). This is mentioned by Ghiggi et al. (2019) themselves in chapter 5.5. But using TWSCs from GRACE (as in this study) actually allows to take such interactions into account as e.g. retention of water in reservoirs is reflected in the basin-scale total water storage. Thus, using datasets like GURN as a proxy for observed runoff can actually make results worse than they actually are.

  Thus, if possible, it would make the study even more comprehensive and convincing if the authors also include a similar evaluation with observed runoff at least for a subset of study basins.

- I find the separation into Köppen-Geiger-classes a bit problematic. Many river basins extend over multiple climate-zones and -classes. Moreover, the climate zone of the headwaters (i.e. where most of the runoff is generated) might be much more important for the characterisation of the basins than the downstream areas. So while it certainly makes sense to somehow categorise and cluster the basins and, hence, try to better understand the water budget closure, I assume that a different "metric" (e.g. drainage area, length of the rivers, relationship between rainfall and runoff, sources of moisture, catchment-scale characteristics (topography, gradients, etc.), advection vs. local evapotranspiration) could give even more insight into the performance of current hydrometeorological datasets. See e.g. https://hydrosheds.org/page/hydroatlas for a quite new approach for characterising river basins. If possible, it would be worth to check if some of these metrics give a clearer picture of the regional performance.

- While I really acknowledge the huge amount of work that the authors have put in their study, I think that the manuscript is only a starting point for future research. The authors claim that a positive NSE could be achieved over 99% of the basins but only if we choose the right combination of datasets. Thus, for an area of 35.5 mio. km^2 (according Table 3), ERA5 Land might not be the most realistic dataset but best dataset in terms of bias cancellation. The reasons for this remain (mostly) unknown. This, however, is not the authors fault but somehow due to the very nature of such global assessments. Future studies and regional applications must therefore use the findings from this study (e.g. from figures A12-A15) as a starting point to further explore strengths and weaknesses of individual datasets across different regions. Only then are we able to see improvements in our global hydrometeorolgocal data sources (as the authors also state in the abstract).

Minor comments:

- The paper is very long and contains a huge amount of quite detailed information! I would hence really urge the authors to reduce the number of pages! As a suggestion, they could put the whole description of methods (water budgets, central differences, filtering of hydrometeorological information, RMSD, NSE, etc.) into the supplementary material as these topics have been presented in many other studies.
- The authors mention that they bring every dataset to a resolution of 0.5°. This, however, could lead to issues for coarser datasets (GLDAS, GPCP, etc.), particularly for smaller catchments. Are there any relationships between the resolution of the input datasets and the performance metrics?
- Page 10, line 223: I guess err_cst are simply anomalies, right? Then, err_cst^2 is simply the temporal variability of total water storage changes. If this is the case, I would not call these *errors* as this sounds misleading.
- Page 10, equation 6: Similarly, err_cyc^2 is just the deviation from the annual cycle and, again, using the term *error* for such anomalies is quite misleading.
- Page 11, line 260: The two consecutive enumerations (i.e. lines 257 - 259 and 260 onwards) look a bit weird... Please add at least one sentence for separating these two parts.
- page 12, line 263: ...is within the confidence interval from GRACE TWSCs.
- Page 13, lines 294 - 295: Do you have an explanation why the performance has improved? Is it due to improvements of the consistency or the performance of the hydrometeorological datasets?
- Page 14, lines 315 - 325: Do you have any explanation why TWSC is too low in the wet and too high during the dry season, respectively? According to Fig. A3, this under- and overestimation seems to be quite systematic.
- Page 15, line 350: Important for what?
- Page 18: line 398: For this analysis, we focus on a subset of 132 basins out of the 189, where an excellent budget closure could be achieved.
- Page 19, line 405: This is a dangerous conclusion as it indicates that two very "bad" datasets can still lead to good water budget closure, if there occurs a cancellation of biases (i.e. right for the wrong reason), right? This would mean that e.g. the datasets in Figure 12, that satisfy a cost lower than 0.1, must not necessarily be realistic datasets but, by combining them with other suitable datasets, only achieve a reasonable water budget closure.
- Page 19, lines 12 - 16 and Figure A7: I did not really understand the clustering approach. What exactly are the authors trying to do here? Do they want to define 13 representative catchments and then identify smaller basins that achieve a similar performance? If this is the case, why did they choose 13 "artificial" clusters instead of using e.g. similar climatic or topographic conditions?

- Page 19, line 430: This is an important statement as GPCP has by far the lowest spatial resolution of 2.5° (around 250km). Claiming that GPCP (approx. 250km) performs similar than ERA5 Land (9km) indicates that resolution does not play a big role, even this is generally assumed by the community (especially over complex terrain). Could you find any relationship between the performance of GPCP and the size of the catchments?
- Page 21, lines 450 - 455: The authors suggest that the discrepancies between the TWSC from water budgets and GRACE are somehow related to overfitting of the CLSM. But there is also a huge temporal shift between the two time-series. Are there any explanations for this?
- Page 21, Figure 12 (and A12-A15).: Is there any meaning of the length of an individual section (or dataset)? And at least on long-term averages, we often assume that P should be equal to ET + R but this is not the case for several clusters. Can you explain why this happens here? Or did I misinterpret the figures?
- Page 22, Figure 13: As the distribution of NSE-values might be highly skewed, wouldn't it make more sense to show the median of the 10 best-performing combinations?
- Pages 21 and 39, Captions for Figures 12 and A11: Why are MSWEP, PGF and GRUN outside the boxes