

Dear Reviewer 2

Many thanks for your valuable comments on our manuscript. We were happy to read that you appreciated the paper. Please find below our replies to the comments and how we will implement them in the revised version of the paper. We used *blue italic font* for the comments and black font for our replies.

Best regards,

Franziska Clerc-Schwarzenbach on behalf of all co-authors

General comments

RC: "As a strong believer in the importance of large sample studies in hydrology, I read the paper with much interest. I found the results very interesting despite being unsurprised by the results. The strengths and drawbacks of ERA5 and its little brother ERA5-Land data are fairly well known, especially when it comes to temperature (very good) and precipitation (good with some issues such as regional biases). Potential evapotranspiration is more of an unknown, and I found its use in Caravan a bit perplexing since it is an unknown quantity. I believe that 'exotic' data from reanalysis should be thoroughly validated before their incorporation into any hydrological study."

We are happy to read that you found the paper interesting. We agree that the results can be considered unsurprising as one cannot expect global reanalysis data to be as good as station-based (or regional) data. However, we think that the impact of these differences on the model results can still be surprising to many users of the Caravan dataset and that the poor potential evapotranspiration data may surprise people who assume that the data in a large-sample dataset has a fair quality, as you also mention.

RC: "My understanding is that potential evapotranspiration is computed using the surface energy balance assuming a crop soil surface, as it was included for irrigation purposes. As such, it is not surprising that catchment scale estimates would be severely overestimated in many cases. See Muñoz-Sabater et al. (2021) for example."

The definition of potential evapotranspiration in ERA5-Land changed in November 2021 (as stated on the website, see the first link below, under "Known issues / Definition of Potential Evaporation (PEV) modified"). While Muñoz-Sabater (2021) indeed write that for the vegetation type crops were assumed and that it was also assumed that there was no soil moisture stress, the ERA5-Land documentation states that the computation of potential evapotranspiration in ERA5-Land assumes an open water surface (i.e., pan evaporation) and that the atmosphere is not affected by it. This difference between ERA5 and ERA5-Land is also stated elsewhere in the documentation (see the first link below, under "Guidelines / Actual and potential evapotranspiration"). It is furthermore confirmed in the variable description where the data can be downloaded (see the second link).

- First link: <https://confluence.ecmwf.int/display/CKB/ERA5-Land%3A+data+documentation>
- Second link: <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land>

We will clarify this difference between ERA5 and ERA5-Land in our revised manuscript.

RC: "Despite calling the results 'unsurprising', I believe this paper makes some very good observations that are useful to the community [...]."

Thank you very much for listing our main findings and summarizing why our manuscript is helpful for the community. It was valuable to see that the points that we wanted to make came across clearly.

Specific comments

RC: "I am not fully sure why this is considered an Opinion paper. To me, the breadth of the research work and analysis clearly qualifies it as a research paper. I did not see much 'opinion' in this paper, as most of the arguments/discussions are results-based. Consequently, I suggest this submission be reclassified as a research paper."

Thank you for sharing your thoughts on the category of this paper. Originally we were planning to write a shorter paper and therefore contacted the editors about writing an opinion paper. However, we agree that the paper has evolved in a more detailed analysis that is beyond a typical opinion paper. We would be happy to change the category, but leave the decision to the editor.

RC: "I think the title does a disservice to the paper. I like the catchy phrase, but in reality, I would think that a majority of hydrologists are not familiar with Camel, and an even smaller number are aware of Caravan. A more generic title referring to large sample datasets and global datasets would be more appropriate for the varied readership of HESS."

The main target group for this paper are potential users of the Caravan dataset that are aware of this dataset and the Camels datasets. Thus, we like to keep these two words in the title to grab the attention of the target group. However, we will change the title to "Large-sample hydrology: A few camels or a whole caravan?" to provide potential readers with some more information on the topic of the paper.

RC: "I believe an additional discussion point should be added regarding the choice of a particular hydrological model. It is well known that some models may be more flexible than others at adapting to biases in input variables such as precipitation and easily scale PET with specific calibration parameters. This is mentioned in the paper, but I believe some other hydrological models may perform better than the one used in this study, and the performance drop mentioned in this study may not be as bad. I certainly would not expect the conclusions of this paper to be any different, but this should be mentioned."

We will add to section 4.3 that it is possible that the effect of the Caravan forcing data is different (either smaller or larger) when a different model is used as the sensitivity to the input data may be model dependent. We also do not think that the results will be very different for a different lumped conceptual model but agree that it would be interesting to look deeper into the various effects of the choice of the input data for different models.

RC: "There should be a mention of the upcoming ERA6 reanalysis. The ERA5 reanalysis used in Caravan will soon be a thing of the past. In addition to improved resolution, ERA6 will have a full overhaul of the model physics, including radiation, which is overestimated in ERA5 and likely part of the PET problem, in addition to the issue discussed above. Based on past history, we can expect a significant performance increase with ERA6. This should be mentioned in the paper. I believe that reanalysis is indeed the future of large-sample hydrology and that merging reanalysis with Deep Learning approaches will produce very high-quality global datasets much sooner than most people think. Already, the merging of deep-learning methods with weather forecasting models promises to revolutionize weather forecasting—exciting times."

Thank you for pointing this out. We will add to section 2 that the ERA6 dataset is currently in development and that a quality increase can be expected. While we could not find any papers about ERA6 that are available already, we agree that it is important to show that comparisons of reanalysis data and other data may lead to other results in the future.

RC: "I would also like to add one important advantage of global datasets based on reanalysis that was not mentioned in the paper: they are easily updated once a new version comes out. In addition, new data is produced in near-real time. Comparatively, datasets relying on observations (e.g., Camel) are much more complex to update (missing data, stations being decommissioned, etc.) and, based on past history, are unlikely to be updated at all, or very infrequently. A dataset such as Caravan will still need to be updated, but the process is much more straightforward."

We will add to section 1 that the possibility to simply update a large-sample dataset based on reanalysis data is an advantage of datasets that use globally available forcing data. We will take up this point again in the conclusions (section 6) as well. However, we also want to stress the point that having datasets that remain the same over longer periods has advantages; mainly, that it allows for a better comparison between different studies using the same data.

RC: "The use of 'significant/ly' should be clarified if it is in the 'statistical' sense from the get-go at line 70. In some cases, it clearly is, but not so much in others."

Thank you for making us aware that it is not always clear whether we refer to a statistical significance when we use this expression or to a considerable difference. Indeed, when using the expression for the first time in line 70, it was unconnected to a statistical test. Therefore, we will change the wording to "considerably". We assume that in Table 3 and the corresponding caption, it is clear that we mean a statistical significance. However, we will add the word "statistical" before "significance". For the two occurrences in paragraph 4.2.3 and the occurrence in paragraph 4.2.4 for which we did not mention the corresponding p-value, we will add the p-values to clarify the statistical significance.

RC: "I would suggest the use of PET instead of E_{pot} , with the former being a lot more common, in my opinion."

Even though we see the wide use of PET, we prefer to use (and keep) E_{pot} in this manuscript. As we are already using a lot of abbreviations (all consisting of several capital letters) for the datasets, we think that it is better to use the variant with a subscript for the potential evapotranspiration to make it more distinct. Furthermore, E_{pot} can be used in the formula for the aridity index, whereas the use of PET in an equation would be mathematically incorrect (as it would equal P times E times T).