Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2020-228-RC2, 2021

Comment on essd-2020-228

Anonymous Referee #2

Referee comment on "CAMELS-AUS: hydrometeorological time series and landscape attributes for 222 catchments in Australia" by Keirnan J. A. Fowler et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2020-228-RC2, 2021

This paper and associated dataset sets out to provide a large sample dataset for Austrlia, specifically one in the CAMELS series, joining others on three other continents.

I think the authors have fully delivered on this aim. CAMELS-AUS is an impressive dataset, rigorously derived as with other CAMELS initiatives and clearly articulated in the paper, which is very well-written. I believe it will be a major asset to the research community in Australia and, as the others point out, for researchers globally given the possibility of analyses across the CAMELS domains.

I like how the paper describes the unique Australian context and rationale, as well as the typical explanation of the dataset. The list of spatial/contextual metadata available across CAMELS-AUS is particularly impressive, and I think the authors have made many pragmatic decisions which will help the user community (e.g. using both major climatic datasets as both are well established).

I have only very minor comments, which will be easy to resolve, I imagine.

Finally, I apologise for the slowness of this review.

Minor Comments:

L32. I *think* I know what you mean here by ‘quality codes’ but on its own this expression quite open to interpretation. You might want to avoid ambiguity, e.g. (‘hydrometric data quality codes/flags’). By the way, the second mention at L94 is OK to me as it is clear of the context.

L48. For me another key benefit is that CAMELS are ‘frozen’, citable open datasets that are not subject to change, ensuring repeatability. Whereas going to agencies means data are always likely to be updated (e.g. reprocessed due to rating changes). There’s no guarantee it will be the same data used between different studies. Of course, some users demand the most up-to-date data ad hence those sources are the most suitable in such cases, whereas for large-sample science studies the consistency and repeatability is key. Maybe just worth a line or two on this benefit?

L80. Minor point, but I’m not sure what constitutes ‘relatively deep’. Rephrase? Is it not

the persistence that is so characteristic, compared to most regions?

L85. Again, super minor, but what do you mean by 'model failure'? I know it's probably described in the reference but I'm interested, as many readers will be – worth a very short extra line/parenthesis about what you mean here?

L103 onwards. This is all very sensible, the decision to use the HRS. However, it's not universal that these CAMELS datsets (or any large sample) should be undisturbed sites. In fact, many researchers are interested in the mix of influences, and in other cases networks are deliberately chosen to represent a range of degrees of influence. E.g. in the UK where the CAMELS dataset includes many disturbed sites as they are a subject of study. There is then a separate 'benchmark' set of undisturbed sites (analogous to the HRS) which can be used and compared against the wider set with many and diverse influences. In short: it's not a necessary condition for a Large Sample datset to be free from influences – that was a decision taken in Australia but not UK, Chile etc. You could note this somewhere. This really isn't a criticism, just a clarification.

L192. Excellent line here, to humble British hydrologists like me…..

L200. I think I follow the process here. To an (admittedly ill-informed) outsider like me this seems like desperate measures to have to scrape this info from the website. I'm not challenging the end result, I am sure its fine, but was there no other way to obtain this from BOM (and if so possibly benefit from possibly more 'official' handover of spatial data that may even be higher res, given likely downsizing for the website?).

L227. Do you mean this information is not provided in CAMELS-AUS or just not provided anywhere? I appreciate it may just be how it is, but it would be good for users to find out more (somewhere) on the gap filling if they want to, so if it is available somewhere a signposting may help.

L267 I imagine this is just propagating the approach in the original paper and I respect that decision if so, but this seems like a potentially misleading approach to me (thinking of how users take datasets and use them at face value). I would have thought it may have made sense to remove these cases.

L288. Agreed, this seems like a sensible choice, to put this decision to users, given how established both datasets are.

L304 – 310. It does seem like there is some uncertainty surrounding timing conventions and you have done the right thing in signposting this. I imagine it will not make too much difference (although I note the Jan Seibert reference) and I agree on putting this to the user rather than processing the data further. However, I just wonder whether this is also clearly highlighted within the dataset, i.e. in the metadata for these fields?

L340. Is there any issue with numbering/sub-section names? Here, in 3.6 there is a reference to the 'following subsections' but 3.6 has none. 3.7 does so I assume just numbering errors.

L416. Given the catchments are all HRS, at face value there's a tension here (given that Fig 3 shows some catchments score highly on this attribute with some yellowish dots implying high degrees of disturbance). I guess these are just two different definitions of 'anthropogenic influence', and this distinction is fine. However it may be worth being really explicit about that at the start of this section to avoid ambiguity and incorrect interpretations (L424 goes on to note the lack of regulation in the dataset but you could be clearer early on here).

L457. Given CAMELS inherits from HRS, this 'first' is not strictly true in a wide sense, even if true in tehe narrower sense of as a dataset specifically for large-sample analysis with associated metadata etc.