Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2021-58-RC1, 2021

Comment on essd-2021-58

Anonymous Referee #1

Referee comment on "GeoDAR: georeferenced global dams and reservoirs dataset for bridging attributes and geolocations" by Jida Wang et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-58-RC1, 2021

Wang et al. describe the creation of a new global dam and reservoir dataset. Overall, the paper is well-written and the methods and findings are on the whole very clear. This dataset will be a valuable contribution to the community and will likely be used by scientists across multiple fields.

The manuscript is incredibly detailed – perhaps slightly too much so – but users of the dataset may be grateful for it to be so well-documented. On the one hand, it is commendable that the authors have put so much effort into this dataset and into ensuring the manuscript thoroughly describes the entire process; on the other hand, the manuscript does get so repetitive in places, particularly in the results, that it is challenging for the reader to digest. It is up to the authors' judgement as to whether or not to shorten the paper, but I would suggest in general thinking about streamlining the paper in places and shortening a few of the most repetitive sections.

Another more general comment is that this paper uses a lot of jargon (particularly geo-matching and geocoding) which are not clearly defined early in the text. Since most readers will not be familiar with geo-matching or geocoding, I suggest clearly defining these terms at some point early in the paper to explain in better detail what exactly they refer to. The methods are eventually pretty clear, and after reading the paper I now understand what these terms refer to, but a lack of understanding of these terms is perhaps likely to confuse the reader at the beginning of the paper.

As for the dataset itself, it appears complete and is easy to visualize and play around with in GIS. The dam/reservoir attributes included for each polygon as well as the readme are very useful and clear. Clearly a lot of thought and careful work has gone into the dataset and it is very well done.

I have several additional, mostly minor, comments listed below.

Specific Comments:

Line 55: It would be helpful here to define what is meant by "attributes" – i.e. reservoir use, storage capacity, dam completion year, etc

Line 107: Change to something like: "Our preference was the former when possible to optimize the georeferencing accuracy"

Line 109: Change "with" to "by"

Line 117-118: The sentence "We acknowledge that although we tried …" is a bit vague. I suggest adding a bit more detail about the challenges associated with duplicate record removal, making the sentence something like: "We acknowledge that owing to the challenge of XXXX, our duplicate removal is not perfect and may have misidentified or missed some duplicate dams"

Line 135: It is unclear to me what you mean by "build associated between new dams supplemented by GRanD and the WRD records"

Figure 2, particularly Figure 2a, is a little hard to interpret. I understand what you are going for here but I find it hard from looking at these Venn diagrams to determine which of these circles represent what is actually included in the GeoDAR datasets. I suggest revising, adding additional description or considering whether this figure is necessary.

Lines 220-231: Can you provide more details about how the QC/QA was performed? How did this analysis lead to determining that 3% of the matched results were matching errors?

Line 247: The phrase "the forward geocoding input the text address of each dam" does not make sense (and is not grammatically correct)

Line 247: Reading this paragraph, I was a bit confused by what is meant by "text address of each dam" and how this can be used to query the longitude and latitude. While this is explained in greater detail and is much clearer in the following paragraph, it made this

first paragraph hard to follow. I suggest perhaps reorganizing this section (i.e. putting some specifics from the following paragraph into this paragraph) and/or providing a short primer, either here or earlier in the text, about "forward" vs. "reverse" geocoding so the reader, who is likely unfamiliar with geocoding, can better understand this rather abstract description.

Line 297: Can you clarify, both here and in the above mention of QA/QC (see above comment) whether all reservoirs were manually QCed, or just a subset?

Line 303: This offset for dams in China is interesting and as you probably know likely has to do with China's GPS shift problem. Perhaps another sentence could be added to explain this in greater detail (i.e. why Google Maps does not work in China like it does in the rest of the world)

Lines 435-440: Was any QA/QC performed for the reservoir/dam matching? What is the likelihood that some dams were incorrectly matched with their reservoir polygons?

Lines 504-505: The sentence starting with "As a result, the average reservoir size decreased…" is a little confusing as I am unsure whether the decrease in mean size is referring to the mean size of reservoirs identified from each data type (i.e. GRanD, HydroLakes) or the mean size of the entire dataset decreasing as the datasets were added in hierarchical order from largest to smallest. I suggest keeping this information in the text but rephrasing to make clearer.

Line 407: Change to the "the retrieved polygons do not always represent the maximum water extents of the reservoirs…"

I don't think Figure 12 is necessary since it is pretty hard to see differences between the two datasets at the global scale. Figure 13 and Figure 14 are much better illustrations of the differences between the datasets.

I'm not sure the section title "Improved spatial details over GRanD" is the right phrasing. I understand what you actually mean but to me this phrasing implies that you have improved the detail of the spatial attributes of GRanD reservoirs, not actually increased the number and decreased the size of resolved reservoirs. I'm not sure exactly what it should be changed to, but I would consider rephrasing. Similarly, throughout the text, I would suggest rephrasing or coming up with another term to replace "enhancing the spatial detail" since this could mean different things to the reader.

I'm not sure what the takeaway from Figure 16 is intended to be, other than perhaps

showing that in general the datasets don't necessarily agree and include different dams/reservoirs? The color scheme and size of the dots also make it very hard to distinguish against the colored background. I suggest either removing this figure entirely or at the very least changing the color scheme.

The section 3.4.2 (pages 33-41) is very long and detailed, particularly when it can perhaps be summarized in one sentence as something like "GeoDAR contains substantially more small dams/reservoirs than GRanD, and therefore while the total capacity of the reservoirs in GeoDAR and GRanD is similar, the main advantage of GeoDAR is that it is more spatially extensive by including many more small dams and reservoirs." It is up to your judgement, of course, but I suggest considering shortening this section – much of the info here is perhaps interesting, but it is very repetitive and therefore hard to read in places.

The conclusion is really well-written and does a very good job of highlighting the advantages and novelty of this dataset. It is a bit repetitive but I think it is appropriate here because it is more of a summary than a conclusion (and that works for this sort of paper).