Earth Syst. Sci. Data Discuss., author comment AC3
https://doi.org/10.5194/essd-2021-58-AC3, 2021

Reply on RC1

Jida Wang et al.

Author comment on "GeoDAR: Georeferenced global dam and reservoir dataset for bridging attributes and geolocations" by Jida Wang et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-58-AC3, 2021

We sincerely appreciate Reviewer 1 for his/her encouraging and constructive comments. These comments helped us clarify the merits and limitations of our dataset, and improve the structure and readability of our paper. Before our point-by-point responses, we provide a list that summarizes the major changes:

- We reorganized part of the manuscript to better streamline the methods and results. The revised "Methods" section starts with a definition and method overview, followed by the subsections elaborating each of the primary methods. The previous lengthy "Results and discussions" section has been broken into several stand-alone sections, including "Production components and usage", "Validation", and "Comparisons with existing global datasets".
- Both reviewers indicated that our methods and results are overwhelmingly detailed. To improve the readability, we have relocated some of the technical triviality to Supplementary Materials and have reduced the redundancy as much as possible. However, we kept a certain amount of detail that we deemed important. Since this is a data description paper, our rationale is to ensure that we have conveyed the principles as such readers understand how our dataset differs from the existing ones and may potentially replicate and improve our dataset.

Our revision also includes several data improvements not requested by the reviewers:

- We have redone the geo-matching for the US by applying the newest version of the US National Inventory of Dams (version 2018, consisting of ~90K records).
- We improved our scripts to better handle the consistency between the names of states/provinces in ICOLD and those in regional inventories and Google Maps.
- We have repeated part of the QC to detect and correct more geocoding errors (such as misplacements in China and omissions in India).
- When we were harmonizing GeoDAR v1.0 with GRanD, we identified about 70 records in GRanD with possible georeferencing errors. These records were excluded from the revised harmonization. We released these problematic GRanD records, as well as our suggested corrections, in the Supplementary Materials for user convenience.
- Meanwhile, we took a deeper stab at building the linkage between WRD and GRanD. These above-mentioned improvements ended up expanding the total number of dams

and reservoirs in our revised product by about 1300.
- We also expanded the validation sample from the previous ~980 dam points to now more than 1400 dam points. The accuracy turned out to be overall consistent.

Author's response to reviewer comments

Anonymous Referee #1

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.

Response: We appreciate the reviewer's encouraging comments and the recognition of the potential value of our dataset.

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.

Response: We thank the reviewer for this constructive comment. Since this is a data description paper, we documented the methods and results in substantial detail, hoping that any user will not only understand the dataset but also be able to replicate or improve the production. On the other hand, we agree with the reviewer that some reorganization of the contents is needed to reduce repetitiveness and improve clarity. In brief, this is what we have done:

- We have simplified the Methods section by (1) starting with "Definition and overview" followed by several subsections that elaborate the principles of primary procedures, and (2) relocating some of the technical triviality for each subsection into Supplementary Materials. This way, users will have a clear sense of how we streamlined the methods without being too overwhelmed by the technical details.
- We have broken the previous lengthy "Results and discussions" into several stand-alone sections, including "Production components and usage", "Validation", and then "Comparisons with existing global datasets". This way, the main deliverables appear more organized, and meanwhile we felt less pressured to have to cut the details we deem necessary.
- We reduced the redundancy as much as possible in the section "Comparisons with existing global datasets". However, we still kept a substantial amount of detail that we considered important. Since this paper is nothing but data description, our rationale is that providing a well-rounded, comprehensive comparison between our product and other existing datasets will greatly benefit the user when he/she is debating on which one to use.
- We relocated some of the discussions about the applications of our dataset to the

conclusion section (now entitled "Summary and applications").

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.

Response: We thank the reviewer for this suggestion. We have considered two options in the revision. One option is to add a section that exclusively defines all jargons in our paper. However, such a section may look too mechanical; particularly, the definitions may appear disconnected from the context. The other option is to introduce each jargon as early as possible when the method develops; nevertheless, this may also obscure some of the terms as the reader may need to search the entire method section for one definition.

As a compromise of both options, we have decided to do the following:

- We merged the previous "Georeferencing rationale" and "Method overview" sections into one section "Definitions and overview" (Section 2.1). This way, readers are clearly informed this is the section where important definitions (such as geo-matching and geocoding) are introduced.
- In this merged section, we provided a concise method overview. This overview streamlined the key procedures along with important definitions (when possible), but excluded technical details (which were elaborated in the follow-up Method sections). When a jargon cannot be fully explained in the overview section, it was explained later but as early as possible. For the latter situation, we tried to provide sufficient context so the introduction of any jargon will not appear too abrupt or confusing.
- For the convenience of the reader, we tried to be as diligent as possible in cross-referring the same jargon to the occurrences in different sections.

We invite the reviewer to look at our revised method section.

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.

Response: We sincerely appreciate this encouraging comment.

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

Response: As suggested, we have added some explanation to define "attributes":

"… ICOLD WRD provides more than 40 attributes (e.g., reservoir storage capacity, dam height, and reservoir purpose)".

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

Response: Thank you. This sentence has been changed as suggested.

Line 109: Change "with" to "by"

Response: Thank you and this has been changed as suggested.

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"

Response: Thank you. As suggested, we have revised this sentence to:

"We acknowledge that owing to the challenges of lacking explicit spatial information and occasional attribute errors in WRD, 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"

Response: We are sorry about this confusing statement. For improved clarity, we have revised it to:

"The harmonization aimed at merging both datasets, removing duplicates between them, and when possible, associating each new dam supplemented by GRanD with the corresponding WRD record."

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.

Response: Thank you for this suggestion. For improved clarity, we have provided the following explanation in the figure caption:

"Boxes indicate final subsets in each GeoDAR version, and the arrows point to the georeferencing sources or methods. Topology of the shapes illustrates logical relations among the data/methods, but sizes of the shape were not drawn to scale of the data volume."

The reasons that we would like to keep these Venn diagrams are: the dams from some of these data sources or methods (circles) overlap with each other, so we believe using the Venn diagrams is perhaps the most visually effective way for readers to understand their relationships and how they contribute to each of the final components (boxes) in our dataset. Yes, I agree that these Venn diagrams can be a little tricky to interpret (although we have tried to optimize the design and clarity), but I think they offer more benefits than confusion and the readers can also refer to Table 1 for more clarification. In general, I hope the reviewer find this figure, now with our expanded caption, useful and clearer.

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?

Response: Thank you for this question. In brief, we treated quality assurance (QA) and quality control (QC) as two separate processes. QA was firstly performed as an automated filter, followed by QC where we manually verified if the result was indeed accurate. Taking the geo-matching process for example, QA ranked all geo-matching results to several QA levels according to the quality of attribute agreements (see agreement scenarios in Supplementary Table S1). If a WRD record was matched to more than one register records, QA selected the match with the best rank. For any match that did not meet the minimum rank, QA filtered it out of the result. This way, each georeferenced WRD record was only matched to the best-ranking register record. The same principle applied to the QA process for geocoding except that the source for geocoding was Google Maps rather than regional registers (Supplementary Table S3; also see Section 2.3). For more technical details about QA, users can refer to our Python scripts for geo-matching and geocoding at https://github.com/jida-wang/georeferencing-ICOLD-dams-and-reservoirs.

QC aimed at manually reassuring the quality of the georeferencing results after the automated QA. To do so, we went through each georeferenced WRD record to examine whether its attributes (such as dam/reservoir name, administrative locations, river name, and for geo-matching, construction year and storage capacity if possible) generally agreed with those of the georeferencing source (i.e., regional record for geo-matching and Google Maps for geocoding). If any georeferenced WRD record showed a major discrepancy with the source, this record was considered to be erroneously georeferenced and thus removed from the final result. Our manual QC removed ~4% error from the QA'ed geo-matching result and ~42% error from the geocoding result.

To address the reviewer's suggestion, we have restructured the last two paragraphs of Section 2.2 to reflect the explanation above and to improve the clarity of how QA and QC were performed separately.

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

Response: We are sorry about this confusion. In our original sentence, "input" was a verb in the past tense. For improved clarity, we have modified this sentence to be:

"The forward geocoding (see Section 2.1 for definition) used the text address of each dam as the input…"

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.

Response: We really appreciate this constructive suggestion. Following the suggestion, we first clarified the definitions of "forward" and "reverse" geocoding in Section 2.2: "Opposite to regular (or "forward") geocoding which converts a nominal location to numeric spatial coordinates, this reverse geocoding converted the spatial coordinates of each dam documented in the register, to a parsed address that contains administrative divisions at consecutive levels."

Then, we reorganized Section 2.3 (originally Section 2.4) by first introducing how the text address was formatted before describing how geocoding was conducted. Following the framework of geocoding, we then described in more detail how the automated QA filtering was performed. We also simplified the original paragraph for text address formatting and located the technical details of it in Supplementary Text and Supplementary Table S1.

We invite the reviewer to see our revised Section 2.3 and the Supplementary Materials.

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?

Response: Yes, we have now clarified in both sections (2.2 for geo-matching and 2.3 for geocoding) that we screened through the entirety of the georeferenced dams as thoroughly as possible during QC.

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)

Response: Thank you for this suggestion. Yes, the offset for dams in China was caused by China GPS shift problem, as the reviewer commented. As suggested, we have added a sentence to explain this:

"Due to China's GPS shift problem, geocoded points in mainland China tended to show a systematic offset of roughly 500 m from their actual dam or reservoir features."

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?

Response: Thank you for this question. Yes, we browsed through the dam-reservoir pairs and found most matches to be accurate. Problems tended to rise when several reservoirs are close to each other (such as the situation of cascade dams). Since our algorithm has no mechanism to distinguish upstream and downstream drainage positions, this situation may lead to a dam assigned to a downstream (or upstream), larger reservoir. We manually corrected these matching errors if seen. For improved accuracy, we are now going through another round of QC on the retrieved reservoirs and will provide an updated version once the manuscript is officially published. For clarity, we also added a sentence at the end of Section 2.5 (Retrieving reservoir boundaries):

"A manual QC was performed on the combined result to confirm that each retrieved reservoir polygon was matched to the correct dam point, and if not, we tried to adjust the association as thoroughly as possible."

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.

Response: Thank you for raising this confusion. The decrease in mean size refers to the mean size of reservoirs identified from each data type. We have clarified this sentence to:

"As a result, the mean reservoir polygon size decreased from 66 km^2 for those identified from GRanD, to 2 km^2 from HydroLAKES and then less than 1 km^2 from the UCLA Circa-2015 Lake Inventory."

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

Response: Thank you and this has been changed as suggested.

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.

Response: Thank you for this comment. Our rationale of including Figure 12 is that it offers some unique aspects that Figure 13 and Figure 14 do not have.

Different from Figure 14 which aggregated the improvement by country, Figure 12 shows the global distribution of GeoDAR storage capacity dam by dam, thus providing the most spatially-explicit view. We believe we need such a figure to present our global dataset. By comparing such a global distribution with that of GRanD side by side, we aim to convey: (1) GeoDAR improved GRanD in the spatial density of dams, (2) most of the added dams in GeoDAR have relatively small storage capacities, and (3) how the improvement varies across space in the most spatially-explicit way possible.

However, we agree with the reviewer that it is difficult to effectively convey the messages above on a global view. So to supplement Figure 12, we decided to include Figure 13

which blows-up a few hotspot regions. So, if the page limit is not a factor, we prefer to keep both figures.

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.

Response: We appreciate this comment. We agree this expression may have a different connotation. To avoid confusion, we have rephrased "improved spatial details" to be "improved spatial density" in most of the cases that we found confusing. We believe "improved spatial density" is less ambiguous because it indicates a greater quantity of dams or reservoirs per unit area, which is exactly what we meant to say.

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.

Response: Yes, as the reviewer said, the main takeaway of Figure 16 is to illustrate that GeoDAR is not a complete replication of GOODD or GRanD, and with different dams and reservoirs introduced, GeoDAR is able to well complement what GOODD and GRanD have covered in different regions of the world. This figure visualizes such a value of our dataset, as well as a possible benefit of combining all these datasets to achieve a better global coverage.

For this reason, we prefer to keep this figure. But as the reviewer kindly suggested, we have adjusted the color scheme by (1) lowering the brightness of the background satellite images, (2) brighten the colors of the dam points, and (3) increase the size of the dam points.

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.

Response: We sincerely appreciate this constructive comment and found the summary of this section (now Section 5.2) from the reviewer very accurate and to-the-point. Although we fully agree that the main idea of this section can be summarized to a couple of sentences, we do genuinely believe that a detailed and multifaceted comparison between GeoDAR and GRanD, like what we presented, will offer the users more benefits than obstacles. In case a user finds this amount of detail disorienting, we made sure to start

every paragraph with a clear takeaway. This way, a user can easily understand what we are comparing and then decide on whether to read or skip this paragraph depending on the user's interest. In general, we streamlined this section by the following takeaways:

- GeoDAR introduced a larger quantity of smaller dams despite a limited increase of total storage capacity.
- The improved spatial density of smaller dams is almost ubiquitous across the continents.
- Although GeoDAR's improvements are widespread, the improvement levels are not spatially uniform.
- Certain regions with a limited increase in dam count show a greater increase in storage, implying that GeoDAR also improved the inventory of large dams.
- Other benefits of GeoDAR include a more compete representation of regulated watersheds and the identification of many smaller reservoir boundaries.
- Finally, GeoDAR improved the quantity of dams for all primary purposes, which will potentially benefit the understanding of reservoir operation rules.

While these takeaways were maintained, we tried to shorten some of the text to reduce unnecessary redundancy. We invite the reviewer to look at the revised Section 5.2.

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).

Response: Thank you very much for this positive comment.