Earth Syst. Sci. Data Discuss., author comment AC4
https://doi.org/10.5194/essd-2021-58-AC4, 2021

Reply on RC2

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-AC4, 2021

We sincerely appreciate Reviewer 2 for his/her 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 #2

Wang et al., describe a new global georeferenced database of dams based on geo matching attributes from the proprietary ICOLD database with publicly available sources to match attributes to spatial locations of dams.

This database complements existing geo-referenced databases as it a) expands on the number of dams for which more attributes such as reservoir storage are available and b) allows for connecting spatial locations of dams with attributes from the ICOLD database. As such I believe this database is a valuable addition to the growing number of global dam datasets (e.g. see www.globaldamwatch.org).

Response: We are grateful for the reviewer's recognition of our data values.

Whilst this database certainly has merits I think it promises more than it delivers. The authors frame the paper as a significant improvement over other dam databases in that it includes more attributes as there are already other global datasets (e.g. GOODD and GROD) that include more dams but lack such attributes. However, the majority of the paper is focused on identifying the spatial location of dams and improvements in quantity and spatial locations of dams. E.g. line 812 "GeoDAR's major improvement lies on the quantity or spatial details of the dams". This framing is understandable given the challenge of linking to a guarded proprietary database. However, this limits the use of the database as only people who have purchased access from ICOLD maybe able to connect the attributes with the spatial location of the dams. Whilst this point is made clear in the conclusions it could be made more clear in abstract and introduction and the overall framing of the paper. It might be worth focusing on potential applications of the dataset.

Response: We very much appreciate this comment. The reviewer is correct that our dataset improved the quantity of georeferenced dams but the access to their attributes is conditional on a purchase from ICOLD. The latter is restricted by the proprietary nature of ICOLD which we bear no responsibility for. The contributions exclusively made from us are 1) freely-accessible dam and reservoir features that improved the spatial density of existing global datasets, and 2) a way to enable the use of ICOLD attributes for more spatially-explicit applications. In other words, the geometric features are what GeoDAR can directly offer, and the access to attribute information is an extended capability of GeoDAR.

To incorporate the reviewer's suggestion, we have further clarified these merits and limitations in the abstract and the introduction.

In the abstract, we have clarified: "GeoDAR does not release the proprietary WRD attributes, but upon individual user requests we may provide assistance in associating GeoDAR spatial features with the WRD attribute information that users have acquired from ICOLD. Despite this limit, GeoDAR with a dam quantity triple that of GRanD, significantly

enhances the spatial details of smaller but more widespread dams and reservoirs, and complements other existing global dam inventories."

In the Introduction, we have restructured the last paragraph into two paragraphs. The ending paragraph clarified the limit of GeoDAR and suggested potential usage.

"For proprietary reasons, neither GeoDAR version releases any WRD attributes. Instead, we provide an option for users if they need to acquire the attributes: upon individual request we may assist the user who has purchased WRD (https://www.icold-cigb.org/GB/world_register/world_register_of_dams.asp) to associate the GeoDAR ID with the ICOLD "international code", through which WRD attributes can be linked to each GeoDAR feature (see Sections 3.3 and 7 for more details). Even without the proprietary WRD attributes, GeoDAR offers one of the most extensive and spatially-resolved global inventory of dams and reservoirs, which may benefit a variety of applications in hydrology, hydropower planning, and ecology."

More discussions about potential applications of GeoDAR are also given in the concluding section (now "Summary and applications").

Overall, it is also not entirely clear to me why there are two versions of the dataset released simultaneously. It seems to me that V1.1 supersedes v1.0 in that it includes more dams and associated reservoirs and the harmonising with GranD is just part of the method. The authors in line 929 also refer to V1.1 as "our end product".

Response: Thank you for this question, and we do take responsibility for this confusion. Following our response to the previous comment, we have provided the reason of including both versions at the end of the last second paragraph in the Introduction section:

"While GeoDAR v1.1 can be considered as a version that supersedes v1.0, the latter was georeferenced independently from GRanD, and we opted to release both versions so that users have the flexibility to choose whichever works better for their cases and potentially improve the harmonization."

As noted by the authors, (line 967) connecting the dam locations with a hydrographic network would enable research into hydrological implications and ecological connectivity. This would greatly enhance the utility of the dataset.

Response: Thank you for echoing the value of this potential application. Although snapping the dam locations to river networks will extend the applications of our data (as we discussed), this was not yet the primary goal of this data paper. We here focus on improving the spatial inventory of global dams and reservoirs, which is more fundamental but still essential to the improvements of water "infrastructure" data. As both reviewers agreed, what we have produced already represents an important contribution, and following this advancement, we will consider a future revision that rectifies this product to global river networks.

As also noted by an earlier reviewer, the paper is very detailed and quite repetitive. I think it could be significantly shortened to make it more readable. In particular, the methods section is very detailed. Whilst this may be useful for some readers, the majority

of readers will not require such extensive detail and could perhaps be referred to supplementary material if more detail is required. The methods section already includes a lot of the numbers later presented in results and discussion while some validation methods get introduced in the results section so some re-organisation would be required.

Response: We are grateful for this constructive comment. Since this comment echoes that of reviewer 1, we here reiterate some of our reasoning and the corresponding changes below.

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. However, we agree that some of the text appears repetitive, and some reorganization and simplification are needed for an improved readability. We summarized what we have revised below:

- We started the Method section with "Definition and overview", which was then followed by the subsections that elaborate the principles of the primary procedures.
- In the subsections for geo-matching and geocoding (Section 2.2 and 2.3), we relocated some of the technical triviality 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 also removed some of the reported numbers in Methods so that they won't appear too repetitive with those in Results. However, we kept some of the intermediate numbers when we believe they are necessary for the clarity of method description.
- We have re-organized the paper so that the Methods section (Section 2) is exclusively dedicated to data production, and the methods for validation are treated separately from data production and are included in the Validation section (now Section 4).
- Following the point above, we have broken the previously lengthy "Results and discussions" into several stand-alone sections, including "Production components and usage", "Validation", and then "Comparisons with existing global datasets".
- 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").

Specific comments:

I'm surprised that only about 60% of dams from GRanD were found in GeoDAR considering GRanD dams tend to be the largest and usually well documented dams and as such I would expect their attributes to be easily found.

Response: Thank you for this important question. First of all, please allow us to clarify (just in case of misunderstanding) that GeoDAR v1.0 does not georeference all records in ICOLD WRD. So, the percentage of GRanD found in GeoDAR v1.0 (which is a georeferenced subset of WRD) will be lower than the percentage of GranD found in the entirety of WRD.

In our initial manuscript submission, the percentage of GRanD found in GeoDAR v1.0 is 64% (i.e., 4691 out of 7320) as the reviewer pointed out, and the percentage of GRanD found in the entire WRD is 85% (i.e., 6209 out of 7320). This means that the harmonization with GRanD helped us match another 1518 WRD records that were not georeferenced in GeoDAR v1.0.

In the revision, we improved the georeferencing scripts and did a deeper search in between WRD and GRanD (please refer to our revision summary at the beginning of the response letter). As a result, the percentage of GRanD in the updated GeoDAR v1.0 increased to 69%, and the percentage of GRanD found in the entire ICOLD WRD reaches 89%. The fraction of GRanD not found in WRD includes only 810 dams (or 11%), which were also appended to GeoDAR v1.1.

The revised result reflects a harmonizing effort that was as thorough as our capability allows at this moment. On a relevant aspect, this is also one of the reasons why we released both GeoDAR v1.0 and v1.1, in case users want to use v1.0 to perform their own harmonization with GRanD with a possible improvement.

In Section 5.3, we acknowledged that some of the remaining 810 dams in GRanD might be documented in WRD, although matching them was probably tricky due to the challenges of attribute inconsistency between the two datasets and the lack of spatial explicitness in WRD.

Given the chance, we would like to mention that GeoDAR v1.1 is not a terminal version of our data. Should additional GRanD dams be found in WRD, we may consider adding these associations in an updated GeoDAR version.

I was wondering if there could be a potential bias in WRD data since this is a volunteered database? Are there any countries not included because they don't contribute to ICOLD?

Response: Thank you for raising this point, and this could well be. For instance, among the 59k original WRD records, more than 23k (about 40%) come from China alone, whereas Russia is only documented with 70 or so dams. This stark contrast indicates the existence of biases among the contributing nations.

To absorb this comment from the reviewer, we added in Line 753 (Section 5.2):

"However, this pattern also reflects the disparities due to several factors, such as a possible bias in WRD (as it is a volunteered dataset and not all member nations contributed equally), the accessibility of regional registers for geo-matching, and geocoding challenges for different countries."

Line 54. "inaccessible" in what sense? I believe many WRD coordinates can be made available at cost. Suggest change to e.g. not freely or publicly available. In particular as the point about public availability is made in line 58 for regional registers.

Response: As suggested, we have change it to "not publically available".

Line 93. We may decrypt? Perhaps link to more detail provided in sections 4 and 5.

Response: We have rephrased this statement to improve the clarity:

"… upon individual request we may assist the user who has purchased WRD (https://www.icold-cigb.org/GB/world_register/world_register_of_dams.asp) to associate the GeoDAR ID with the ICOLD "international code", through which WRD attributes can be linked to each GeoDAR feature (see Sections 3.3 and 7 for more details)."

Line 98. How is it possible that about 1/3 of the WRD dams (v1.1) capture a similar total storage capacity as the full WRD inventory of ~60,000 dams. Is this because the remaining ~40k dams in WRD are non-reservoir dams? Please explain this in this section. Also would be good to provide the total storage in WRD here (which is only provided in line 138)

Response: Thank you for this question. We see three major reasons.

First, the original storage capacity values in ICOLD WRD have occasional voids or unit errors (sometimes underestimated by a factor of 1000, or for some of the US dams, the values in acre feet instead of thousand cubic meters), whereas in GeoDAR v1.1 we overwrote the original WRD capacity values by those of GRanD (when available). This correction could lead to an increased capacity in GeoDAR v1.1. Because of this, we later in Section 5.2 replaced the original capacity values of WRD that overlaps GRanD by the capacity values of GRanD, and then used the updated WRD capacities for comparison with GeoDAR v1.1 (this way, the comparison will be more apple to apple). Please refer to Section 5.2 for details.

Second, GeoDAR v1.1 absorbed the 810 GRanD dams we were unable to find in WRD. If these 810 dams are indeed not included by WRD, they may also result in a total storage capacity of GeoDAR that is similar to that of the full WRD.

Third, our harmonization between GeoDAR v1.0 and GRanD in the last round included some duplication errors, leading to an amplified storage in GeoDAR. Such duplication errors have been eliminated as thoroughly as possible in the revision.

Now in our revised GeoDAR 1.1, the total storage capacity is 7297 km^3, which is below either the total capacity based on original WRD values (7334 km^3) or the total capacity based on GRanD-adjusted WRD values (7642 km^3). This appears more reasonable. The difference between GeoDAR and WRD capacities (up to ~350 km^3) could be explained by the remaining ~60% of the WRD dams that were not georeferenced. These dams are mostly small, so it is not surprising that their accumulative capacity appears marginal.

Line 118: "We acknowledge…" I suggest rephrasing this sentence to something like: "Whilst we have made every endeavour to remove duplicates, we acknowledge that some duplicates may remain in the dataset"

Response: Thank you. This suggestion concurs with that of Reviewer 1. To combine both suggestions, we have rephrased 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 138. 7388 km3 in original WRD. Is this the figure for all WRD dams or for the cleaned version of 56,783 dams?

Response: This value is the total storage capacity based on the original attribute values of the WRD dams after duplicate removal. In the revision, we performed another round of duplicate removal and concluded a total of 56,850 unique dams in WRD with a total capacity of 7334 km^3.

For improved clarity, we added a sentence in Section 2.1:

"Unless otherwise described, the ICOLD WRD mentioned in the following text refers to the version after duplicate removal."

Line 139: I don't think the Venn diagrams are very clear. Not sure if they are even needed as the text explains the process. A simple flowchart might be easier to understand.

Response: We appreciate this comment, which echoes Reviewer 1 as well. As we responded previously, the reasons that we included the Venn diagrams (rather than a flow chart) 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 topological relationships and how they contribute to each of the final components (boxes) in our dataset.

We agree that these Venn diagrams can be a little tricky to interpret although we tried to keep it simple and clear. But for the reasons above, we think the diagrams offer more benefits than confusion and the readers can also refer to Table 1 for more clarification.

For improved clarity, we have provided the following explanation in the figure caption:

"Boxes indate 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."

We hope the reviewer finds our Venn diagrams, particularly with the revised caption, more acceptable.

Line 239: "rest parts of the world" is a strange phrase

Response: We have changed this to "the other regions of the world".

Line 322: ICOLD storage capacity erroneous. See earlier comment (line 98) on ICOLD reported storage. This can also explain the discrepancy. Note that Mulligan et al (2020) also note erroneous reporting of catchments in ICOLD.

Response: Thank you. We have also cited Mulligan et al. (2020) in this sentence to acknowledge that occasional capacity errors in ICOLD were similarly noticed in other literature and to further backup our statement.

Line 525-562, section 3.2 in results and discussion seems to introduce more methods on validation. This should be moved to methods.

Response: Thank you for this comment. We agree that the original section 3.2 included both validation methods and validation results. This reflected our perception that data validation was not necessarily part of data generation and that we intended to only include the methods related to data generation in the Methods section. This arrangement also reflected our hesitation to group everything after methods into a single "Results" section.

After deliberations, we decided to restructure the previous lengthy "Results" section into a few stand-alone sections. They are "Product components and structure", "Validation", and "Comparisons with existing global datasets". These sections are relatively independent, and arguably not always about "results" (they are also about discussions and applications). So we believe it is more reasonable to reorganize them as separate sections, rather than lumping them all into a single "Results" section.

This way, the "Methods" section now only includes the methodology directly related to data generation, including pre-processing, georeferencing, QA/QC, harmonization, and reservoir retrieval. The product validation, including the methods of sample collection and validation, are now designated to the "Validation" section. We believe this adjusted structure is clearer to the general readers.

Line 797: I find the term (global) capacity improvement a bit confusing. I guess what is meant is a higher reporting of total dam storage capacity by country or globally which is hardly surprising given that more dams and reservoirs are included.

Response: Thank you for pointing out this confusing statement. We agree and have rephrased this expression to "global storage capacity increase", to be comparable to our expression in the previous sentence ("global dam count increase").

Technical corrections:

Numbers in some cases use thousand separator (e.g. 7,163 line 238) but not in others. Please be consistent throughout

Response: Thank you for this meticulous review. For consistency, we have now avoided using the thousand separator for any number under 10,000. But we will make necessary changes upon the request of the publisher.

Line 163. "NID records were accessed"

Response: Thank you. This sentence was deleted as it is no longer necessary. In the revised data we have used a more updated version of USNID (version 2018). Please see the revision summary at the beginning of the response letter.

Line 223 "This led to a conservative success rates"

Response: Thank you. We have corrected “rates” to “rate”.

Line 258: “this process was repeated”

Response: This has been corrected.

Line 570 “We believed”

Response: We have deleted “We believed” in this sentence.

Line 700: Mulligan et al (2020)

Response: Sorry about this typo, and we have corrected “2000” to “2020”.

Line 744: GeoDAR

Response: Thank you. We have corrected this typo.

References: line 39 Doll should be Döll, line 42 Vorosmarty should be: Vörösmarty and there may be others.

Response: We are sorry about leaving out the umlaut. We have corrected it throughout the revised manuscript.