GeoDAR: Georeferenced global dams and reservoirs dataset for bridging attributes and geolocations

Jida Wang[1], Blake A. Walter[1], Fangfang Yao[2], Chunqiao Song[3], Meng Ding[1], Abu S. Maroof[1], Jingying Zhu[3], Chenyu Fan[3], Jordan M. McAlister[4], Safat Sikder[1], Yongwei Sheng[5], George H. Allen[6], Jean-François Crétaux[7], and Yoshihide Wada[8]

[1]Department of Geography and Geospatial Sciences, Kansas State University, Manhattan, Kansas, USA
[2]Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, Boulder, Colorado
[3]Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing, China
[4]Department of Geography, Oklahoma State University, Stillwater, Oklahoma, USA
[5]Department of Geography, University of California, Los Angeles (UCLA), Los Angeles, California, USA
[6]Department of Geography, Texas A&M University, College Station, Texas, USA
[7]Laboratoire d'Études en Géophysique et Océanographie Spatiales (LEGOS), Centre National d'Études Spatiales (CNES), Toulouse, France
[8]International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria

Correspondence to: Jida Wang (jidawang@ksu.edu)

Abstract. Dams and reservoirs are among the most widespread human-made infrastructure on Earth. Despite their societal and environmental significance, spatial inventories of dams and reservoirs, even for the large ones, are insufficient. A dilemma of the existing georeferenced dam datasets is the polarized focus on either dam quantity and spatial coverage (e.g., GOODD) or detailed attributes for a limited dam quantity or region (e.g., GRanD and national inventories). One of the most comprehensive datasets, the World Register of Dams (WRD) maintained by the International Commission on Large Dams (ICOLD), documents nearly 60,000 dams with an extensive suite of attributes. Unfortunately, the WRD records provide no geographic coordinates, limiting the benefits of their attributes for spatially explicit applications. To bridge the gap between attribute accessibility and spatial explicitness, we introduce the Georeferenced global Dams And Reservoirs (GeoDAR) dataset, created by utilizing online geocoding API and multi-source inventories. We release GeoDAR in two successive versions (v1.0 and v1.1) at https://doi.org/10.5281/zenodo.6163413. GeoDAR v1.0 holds 22,560 dam points georeferenced from WRD, whereas v1.1 consists of a) 24,783 dam points after a harmonization between GeoDAR v1.0 and GRanD and b) 21,515 reservoir polygons retrieved from high-resolution water masks. Due to geocoding challenges, GeoDAR spatially resolved ~40% of the records in WRD which, however, comprise over 90% of the total reservoir area, catchment area, and reservoir storage capacity. 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. Along with its extended attribute accessibility, GeoDAR is expected to benefit a broad range of applications in hydrologic modelling, water resource management, ecosystem health, and energy planning.

1 Introduction

Since around the 1950s, the world has seen an unprecedented boom in large dam construction as a response to the ever-growing human demands for water and energy (Chao et al., 2008; Wada et al., 2017). Today, dams and their impounded reservoirs are ubiquitous across many global basins, providing multiple services that range from hydropower and flood control to water supply and navigation (Belletti et al., 2020; Biemans et al., 2011; Boulange et al., 2021; Döll et al., 2009; Grill et al., 2019). These benefits were, however, often gained at the costs of fragmenting river systems, submerging arable lands, displacing population, and disturbing climate regimes (Carpenter et al., 2011; Cretaux et al., 2015; Degu et al., 2011; Grill et al., 2019; Latrubesse et al., 2017; Nilsson and Berggren, 2000; Tilt et al., 2009; Vörösmarty et al., 2003; Wang et al., 2017).

Despite such environmental and societal significance, our spatial inventory of global dams and reservoirs, even for the large ones (such as those with a surface area >1 km^2), has been insufficient. We still lack a thorough and authoritative dataset that documents both geographic coordinates (latitude and longitude) and standard attributes (e.g., purpose, reservoir storage capacity, and hydropower capacity) of the existing large dams. One of the most comprehensive datasets, the World Register of Dams (WRD), is regularly updated by the International Commission on Large Dams (ICOLD; https://www.icold-cigb.org), a non-governmental organization dedicated to the global sharing of professional dam/reservoir information. The recent version of ICOLD WRD documents nearly 60,000 “large” dams, defined as those with a wall higher than 15 m or between 5 to 15 m but with a reservoir storage greater than 3 million m^3 (mcm). These WRD records are considered to be “complete” to the extent of contributions from willing nations and water authorities (Wada et al., 2017).

While ICOLD WRD provides more than 40 attributes (e.g., reservoir storage capacity, dam height, and reservoir purpose), the dam locations are, unfortunately, either not georeferenced or publically available. Despite the availability of many essential attributes, missing geographic coordinates has severely limited the applications of WRD, including for hydrological modelling and hydropower planning (Yassin et al., 2019) which require the dam records to be spatially explicit. This dilemma may be partially resolved by using georeferenced regional registers such as the United States National Inventory of Dams (US NID; https://nid.sec.usace.army.mil). Nevertheless, such regional registers are not always publicly available, especially in developing nations where dam construction is still booming (Zarfl et al., 2015).

Other global dam and reservoir datasets that are georeferenced, however, often lack essential attributes. An example is the recently published GlObal geOreferenced Database of Dams (GOODD V1) (Mulligan et al., 2020), which contains 38,667 dam points digitized from Google Earth imagery and their associated catchments delineated from digital elevation models (DEMs). Despite this dam quantity, GOODD provides no other attribute information. Another inventory, the Global River Obstruction Database (GROD) (Whittemore et al., 2020; Yang et al. 2022), located more than 30,500 flow obstructions along rivers wider than 30 m as mapped in the Global River Width from Landsat (GRWL) database (Allen and Pavelsky,

2018). The current attributes are limited to obstruction types such as locks, weirs, and multiple types of dams. In addition, GROD is tailored for the forthcoming Surface Water and Ocean Topography (SWOT) satellite mission which is designed to observe river reaches wider than 50–100 m (Biancamaria et al., 2016). While these rivers are sufficiently captured by GRWL, the obstruction infrastructure identified along the river mask in GRWL excludes many large dams on rivers narrower than 30 m. In the US, for instance, there are at least 5,170 NID-registered dams higher than 15 m (i.e., large dams according to ICOLD criteria), but less than 8% of these dams intersect with GRWL (i.e., located on rivers wider than 30 m).

Among the few global dam/reservoir datasets that provide both georeferenced locations and essential attributes, are the United Nations Food and Agricultural Organization (FAO) AQUASTAT (Li et al., 2011) and the Global Reservoir and Dam database (GRanD) (Lehner et al., 2011). GRanD was constructed by harmonizing AQUASTAT and a wide range of regional gazetteers and inventories. Its latest version, v1.3, contains 7,320 dams as well as their reservoir boundaries and approximately 50 attributes, with a cumulative storage capacity of 6,881 km^3. Since its publication, GRanD has been applied extensively by a variety of studies, although its focus is on the world's largest dams (e.g., >0.1 km^3) and its quantity (7,320 dams) is a fraction of the 59,000 dams documented in WRD. A spatially resolved inclusion of additional large dams, such as those in compliance with the ICOLD definition, has been increasingly desired by the hydrology community and encouraged by growing collaborations from multiple disciplines such as biogeochemistry, ecology, energy planning, and infrastructure managements (Belletti et al., 2020; Boulange et al., 2021; Grill et al., 2019; Lin et al., 2019; Wada et al., 2017).

Here, we present the initial versions of the Georeferenced global Dams And Reservoirs dataset, or GeoDAR. We built GeoDAR by leveraging multi-source dam and reservoir inventories and the Google Maps geocoding API. Our goal is to tackle the limitations of existing datasets by offering a dam inventory that is both spatially resolved and has an extended ability to access important attributes. As summarized in Table 1, our GeoDAR product includes two successive versions. GeoDAR v1.0 is essentially a georeferenced subset of ICOLD WRD. It contains 22,560 dam points, each indexed by an identifier (ID) that is associated with a WRD record, allowing for the potential retrieval of all its 40+ proprietary attributes from ICOLD. GeoDAR v1.1 consists of a) nearly 25,000 dam points which harmonized v1.0 and GRanD for an expanded inclusion of the largest dams, and b) the reservoir boundaries for most (87%) of the dam points based on a one-to-one relationship between dams and reservoirs. Due to geocoding challenges, GeoDAR v1.0 spatially resolved about 40% of the individual dams in WRD. However, these georeferenced locations were quality controlled, and after the harmonization with GRanD, v1.1 captures a total storage capacity of 7,384 km^3, a magnitude comparable to the full storage capacity of WRD. While GeoDAR v1.1 can be considered as a version that supersedes v1.0, the latter was, in principle, georeferenced independently from GRanD. We opted to release both versions so users have the flexibility to decide whichever works better for their cases and potentially improve the harmonization.

For proprietary reasons, neither GeoDAR version releases any WRD attributes. Instead, we offer 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.

Table 1. GeoDAR product versions and components

Version	Description	Component	Acquisition sources/methods	Count	Storage capacity (km^3)	Reservoir polygon area (km^2)
v1.0	Georeferenced ICOLD	Dam points	Geo-matched via regional registers	13,149	1,308.2	---
			Geocoded via Google Maps API	9,278	1,232.4	---
			Supplemented by Wada et al. (2017)	133	3,900.0	---
			Total	**22,560**	**6,440.6**	---
v1.1	Harmonized ICOLD and GRanD	Dam points	GeoDAR v1.0 alone (excluding overlap with GRanD v1.3)	17,480	507.2	---
			GRanD v1.3 and GeoDAR 1.0	5,080	6,006.0	---
			GRanD v1.3 and other ICOLD	1,414	603.0	---
			GRanD v1.3 alone	809	267.7	---
			Total	**24,783**	**7383.8**	---
		Reservoir polygons	GRanD v1.3 reservoirs	7,120	6,717.7	446,525.2
			HydroLAKES v1.0	7,184	259.8	13,661.9
			UCLA Circa 2015 Lakes	7,211	238.5	36,126.6
			Total	**21,515**	**7,216.1**	**496,313.8**

2 Methods

2.1 Definitions and overview

We aim to georeference (i.e., acquire the latitude and longitude of) each dam listed in ICOLD WRD, by using the nominal location (e.g., a descriptive address for a dam or reservoir) available in the WRD attributes. Examples of the attributes that are important for georeferencing include the names of the dam and reservoir, the administrative divisions the dam is affiliated with, and the name of the impounded river. Using such attribute information, spatial coordinates of a dam may be either a) queried from an existing register or inventory where dam records were already georeferenced and verified, or b) estimated through a geocoding service that can convert nominal locations to numeric spatial coordinates. Our preference was the former when possible to optimize the georeferencing accuracy.

The schematic procedure of GeoDAR production is illustrated in Fig. 1. We started by removing duplicate records from the 59,071 dams listed in the original ICOLD WRD (accessed in March 2019). Here "duplicates" are defined as the dams that are either a) repeatedly recorded with identical (or highly similar) attribute information or b) different dam structures but associated with the same reservoir. Examples of the second scenario include a reservoir's primary and secondary/saddle dams such as the Boonton Dam and its associated Parsippany Dike (40.884° N, 74.408° W) in New Jersey and multiple controls for one reservoir such as Veersedam and Zandkreekdam for Veerse Meer (51.549° N, 3.678° E) in the Netherlands. Although "duplicates" in this scenario refer to different dam bodies, including them could lead to double or multiple counting of the storage capacity of the same reservoir, and similar to the production of GRanD, our goal was to link one reservoir to one dam (if possible). After removing the identified duplicates, the cleaned WRD contains 56,815 unique dams/reservoirs. These dams/reservoirs have an accumulative storage capacity of 7,328 km^3 based on the original WRD attribute values (which occasionally are missing or have unit errors) or 7,720 km^3 after replacement/correction by Wada et al. (2017) and GRanD (see Section 2.4). Unless otherwise described, the ICOLD WRD mentioned in the following text refers to the version after duplicate removal. 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.

We then compared the unique ICOLD WRD records against a collection of georeferenced dam registers we acquired from regional water authorities and agencies. When the attribute information of a WRD dam matched that in a regional register, the spatial coordinates from the latter were "borrowed" by the WRD record. We term this process "geo-matching", which resulted in the georeferencing of 13,190 WRD dams. For the remaining dams in WRD, we applied the alternative approach "geocoding", which transforms a nominal location (such as the dam or reservoir address formulated by ICOLD attribute information) to a pair of spatial coordinates. The tool we used to implement geocoding was the Google Maps geocoding API (http://developers.google.com/maps). The geocoding process successfully retrieved the spatial coordinates of another 9,338 WRD dams. The combined output from both geo-matching and geocoding were next collated with the spatial coordinates and reservoir storage capacities of 133 WRD dams larger than 10 km^3 as documented in Wada et al. (2017). These processes resulted in GeoDAR v1.0, a total of 22,560 georeferenced WRD dam points with an accumulative storage capacity of 6,441 km^3 (accounting for more than 80% of that in ICOLD WRD). The Venn diagram in Fig. 2a provides an overview of the logical relations among the georeferencing sources and methods for GeoDAR v1.0.

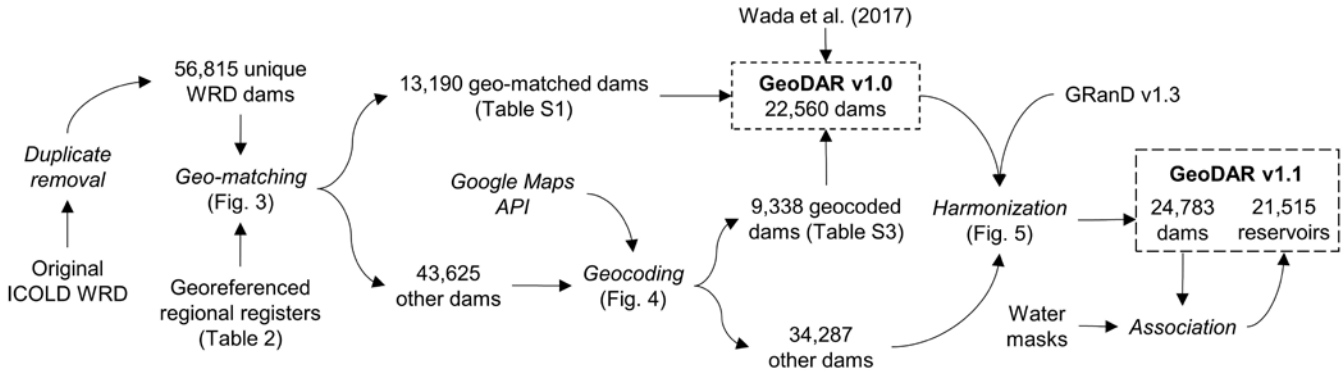


Figure 1. Schematic flowchart of GeoDAR production. Text in roman indicates applied or produced datasets, and text in italics indicates methods or procedures.

To further improve our spatial inventory of the world's largest dams, we performed a harmonization between the dam points in GeoDAR v1.0 and GRanD v1.3. The harmonization aimed at merging both datasets, removing duplicates (overlap) between them, and when possible, associating new dams supplemented by GRanD with the corresponding WRD records. This process identified another 2,223 dam points, including 1,414 associated with WRD but not georeferenced in GeoDAR v1.0. With removal of duplicates, this harmonization led to a total number of 24,783 georeferenced dam points, with an accumulative storage capacity of 7,384 km^3. An overview of this harmonization process is illustrated by the Venn diagram in Fig. 2b. Finally, the reservoir polygons for each of the georeferenced dams were retrieved as thoroughly as possible from three global water body datasets: GRanD v1.3 reservoirs (Lehner et al., 2011), HydroLAKES v1.0 (Messager et al., 2016), and the Landsat-based UCLA Circa 2015 Lake Inventory (Sheng et al., 2016). These nearly 25,000 dam points and their associated reservoir polygons constitute GeoDAR v1.1. Details of production processes and their Quality Assurance and Quality Control (QA/QC) are included in the following method sections.

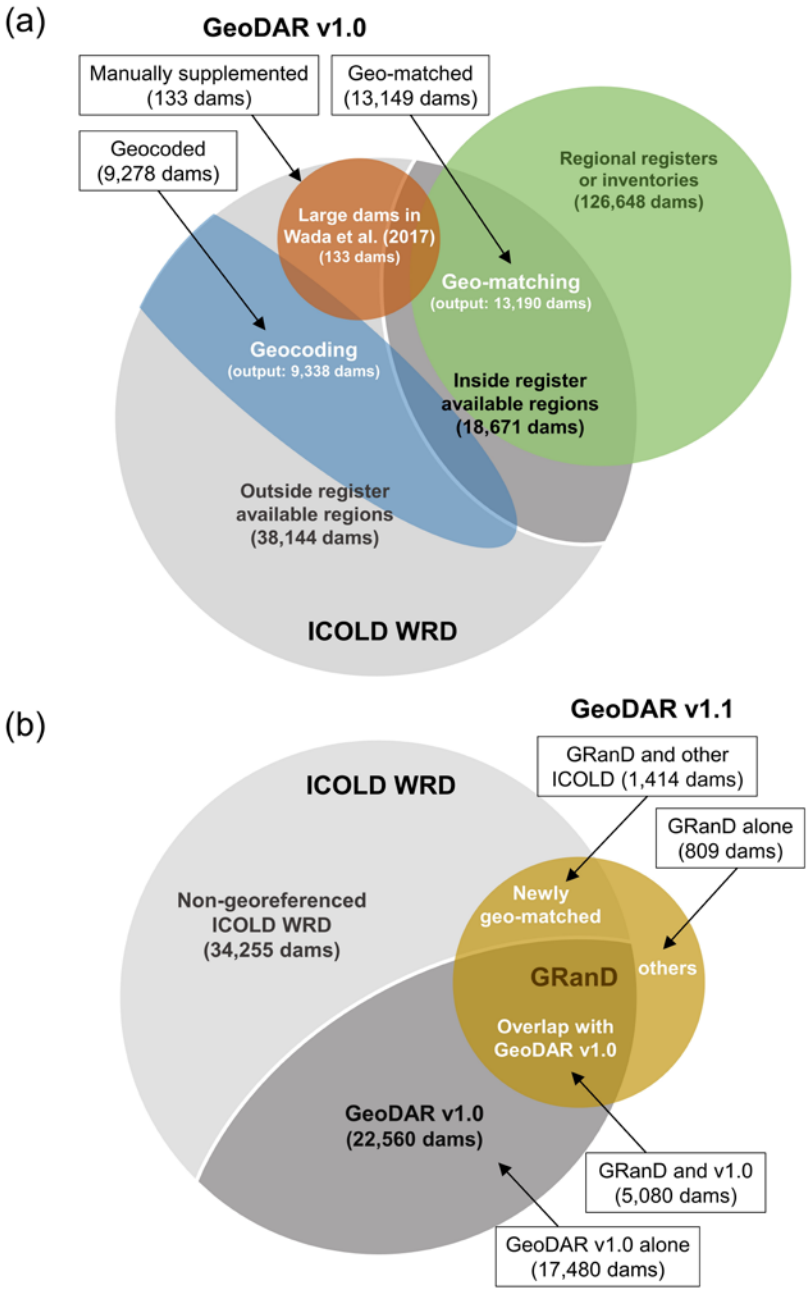


Figure 2. Venn diagrams illustrating the logical relations among georeferencing data sources and methods for GeoDAR. (a) GeoDAR v1.0 and (b) GeoDAR v1.1 (dams only). Boxes indicate final subsets in each GeoDAR version, and the arrows point to the georeferencing sources or methods. Topology of the shapes illustrates logical relations among the data/methods, but sizes of the shape were not drawn to scale of the data volume.

2.2 Geo-matching regional registers

The ICOLD WRD was a joint contribution from more than 100 member nations, some of which also release detailed and publicly accessible dam registers that have been georeferenced. These regional/local registers, with reliable spatial coordinates already provided for each dam, were our preferred sources for georeferencing WRD. Since this type of register is not available for most countries, we searched multiple water authority and project websites, and collected seven georeferenced regional registers or inventories that are open access. Their names, sources, and numbers of documented dams are summarized in Table 2.

Table 2. Regional registers or inventories for geo-matching and the validation of geocoding.

Region	Register/Source	Dam count		
		Regional register	ICOLD WRD	Geo-matched
Geo-matching				
Brazil	RSB (SNISB, 2017)	23,630	1,345	668 (50%)
Cambodia	ODC (2015)	73	7	3 (43%)
Canada	CanVec (NRC, 2017)	843	648	435 (67%)
Europe	MARS (2017)	5,043	6,671	3,981 (60%)
Myanmar	ODM (2018)	254	33	14 (42%)
South Africa	LRD (DWS, 2019)	5,592	1,105	842 (76%)
United States	NID (USACE, 2018)	91,213	8,862	7,247 (82%)
Total		126,648	18,671	13,190 (71%)
Geocoding validation				
China (mainland)	NPCGIS (accessed 2021)	Not counted	23,747	---
India	NRLD (2019)	5,723	5,074	---
Japan	JDF (accessed 2021)	2,349	3,089	---

Register/source acronyms: Relatório de Segurança de Barragens (RSB, Dams Safety Report of Brazil), Open Development Cambodia (ODC), Managing Aquatic ecosystems and water Resources under multiple Stress project (MARS), Open Development Myanmar (ODM), List of Registered Dams (LRD) of South Africa, National Inventory of Dams (NID) of US, National Platform for Common Geospatial Information Services (NPCGIS) of China, National Register of Large Dams (NRLD) of India, and Japan Dam Foundation (JDF). Regional inventories were collected with partial reference to the Global Dam Watch project (http://globaldamwatch.org). Dam numbers for regional registers are based on the records with valid geographic coordinates, and numbers for ICOLD WRD are based on the records after duplicate removal. See full registers, references, and download links in the reference list.

These seven registers/inventories cover Brazil, Canada, the United States, 31 European countries (including part of Russia), South Africa, and part of Southeast Asia (Cambodia and Myanmar), with a total dam count of more than 126,000. Besides

spatial coordinates, each of these registers also provides attributes for their documented dams, which were required by the geo-matching process. While other dam inventories could be available, our geo-matching effort for GeoDAR v1.0 was focused on these collected ones. However, we referred to additional registers from China, India, and Japan (Table 2) for the validation of our WRD geocoding (see Validation). For these additional regional registers, it was either inconvenient to bulk-download the dam records, or we were legally restricted from releasing their dam coordinates. Therefore, we only used these registers for the purpose of validation.

The procedure of geo-matching is illustrated in Fig. 3. Given each regional register, our goal was to find its matching records from the subset of ICOLD WRD for the same region, by cross-checking value similarities for several key attributes between the two datasets. On one hand, the compared attributes must be mutually available in both datasets. On the other hand, the attributes should cover various themes so that in combination, they are able to disambiguate records that represent different dams but may coincide in certain attributes. Taking both requirements into account, the key attributes used include the dam and reservoir names, multiple levels of administrative/political divisions for the dam, and the dam's completion year. The river on which the dam was constructed was also considered for all regions except Cambodia as the register does not contain such an attribute. For each of the key attributes, we considered values in WRD and the regional register agreeing with each other if the similarity score between the value sequences exceeded about 85% (meaning that there are more than 8 pairs of identical elements, with consideration of their orders, between two 10-character sequences). This similarity threshold tolerated minor variations in spelling that often occur among different data sources. If an agreement was not reached between the two full sequences (e.g., "Maharashtra Pradesh" and "Maharashtra"), the similarity was then tested between the main subsets of the sequences in order to increase the matching success.

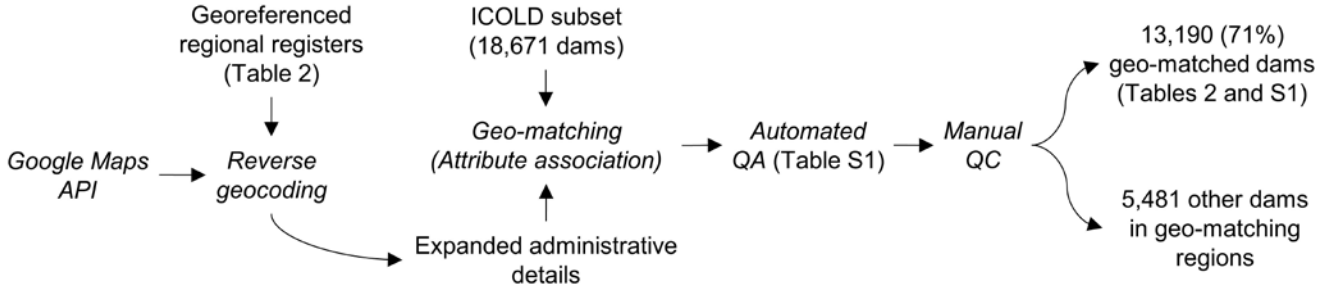


Figure 3. Schematic procedure of geo-matching regional registers. Text in roman indicates applied or produced datasets, and text in italics indicates methods or procedures.

One of the geo-matching challenges was that the levels of political/administrative divisions are not always comparable or consistent between WRD and the regional registers. In WRD, the divisions were provided at the levels of country, state/province, and the nearest town/city, which are inconsistent with some of the registers. For example, the register for

Brazil (Dams Safety Report in 2017) provides the finest division at the county level, whereas the European inventory (from the MARS (Managing Aquatic ecosystems and water Resources under multiple Stress) project) documents no divisions below the national level. To improve the feasibility in division comparison, we performed a "reverse geocoding" for each georeferenced regional register using the Google Maps geocoding API. 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. These multi-level divisions and subdivisions were appended to the original regional registers (Fig. 3), thus enabling a more flexible and complete comparison with the WRD attributes and thus an increased success rate of geo-matching.

We considered a WRD record matched with a regional record if their agreements on the key attributes warranted a reasonable confidence that the two are the same dam. In principle, a high confidence would require a unanimous agreement on all key attributes. However, this ideal scenario was often unnecessary and sometimes impossible. One of the reasons is that the key attributes do not always have valid values. In WRD, for instance, the values of "nearest town" for nearly all (>99%) US dams are null. While this attribute is valid for most other dams, the nearest town/city in WRD is not necessarily the division that administrates or contains the dam as is the case in the township in some regional registers. Another reason is that our collected multi-source datasets were not collated by a universal standard. As a result, inherent discrepancies of the attribute definitions and/or values may exist among the datasets. One example is the dam's "completion year", which could be ambiguous between the year when the dam construction was concluded and the year when the dam operation was initiated or commissioned. These two definitions do not necessarily lead to the same year. To address such inconsistencies, we defined a baseline scenario that required any pair of matched WRD and regional records to agree on the following:

- Dam or reservoir name,
- Country, state/province if values are valid, and
- At a minimum, (a) either completion year or river if the town/city values disagree or are invalid, or (b) town/city when completion years and rivers do not both disagree.

In compliance with this baseline, we implemented an automated QA to filter out any matching errors and optimize the matching accuracy for each WRD record. In brief, any match that did not meet the baseline scenario was removed, and the remaining geo-matched pairs were ranked to three discrete QA levels (M1, M2, and M3) according to the quality of attribute agreements (see definitions in Supplementary Table S1). As the QA rank increases (from M3 to M1), agreements on the key attributes improved from the baseline to the ideal scenario (i.e., a unanimous agreement). If a WRD record was matched to multiple records in the regional register, the QA selected the match with the best rank. This way, each georeferenced WRD record was only matched to the best-ranking regional record. Users may refer to the provided QA ranks as a measure of the general reliability of each geo-matched location. It is worth noting that our geo-matching purpose was to acquire the spatial coordinates of any matched WRD record from the regional register, rather than collating or correcting any existing attribute

values. In other words, some of the WRD and regional records may actually refer to the same dams but were matched unsuccessfully due to major discrepancies between their attribute values. This led to a conservative success rate in our automated geo-matching. More technical details about QA are given in our Python scripts at https://github.com/surf-hydro/georeferencing-ICOLD-dams-and-reservoirs.

Following the automated QA, we performed a manual QC to reassure the accuracy of the geo-matching results. We went through each geo-matched WRD record to examine whether its attributes (e.g., dam/reservoir name, administrative locations, river name, construction year, and storage capacity) indeed agreed with those of the regional source. If an evident discrepancy was identified, the "match" was removed or corrected in the final product. Although we made every endeavour to be as rigorous as possible, remnant matching errors are still possible due to the challenges of incompleteness and intrinsic errors in the attribute information (refer to Section 4 for accuracies). For occasional cases that a dam was matched correctly to the register attributes but misplaced due to poorer quality of the spatial coordinates in the register, we tried to adjust or, if possible, correct the register's spatial coordinates using the best possible resources (such as Google Maps and other open-source documents). If we were unable to observe any water infrastructure at the location of a correct match, we took a conservative action and removed the match. We admit that this might mistakenly delete some of the structures (e.g., small run-of-the-river hydropower stations, weirs, and diversions) that are too small to be visible from Google Map imagery. Our manual QC identified ~4% error in the geo-matched WRD records, most of which came from QA rank M3. After removing these errors, the geo-matching process concluded with a total of 13,190 WRD records georeferenced (Fig. 3), including 3,238, 6,987, and 2,965 for QA ranks M1, M2, and M3, respectively (Supplementary Table S1). The success rate, i.e., the number of geo-matched dams as a percentage of the number of WRD records, varies from about 40% in Southeast Asia to about 80% in South Africa and US (Table 2), with an overall success of 71% in all geo-matched regions (Fig. 3).

2.3 Geocoding via Google Maps

The subset of ICOLD WRD that was not geo-matched includes the remaining 5,481 (29%) dams in the geo-matched regions and the entire 38,144 dams in the other regions of the world (Fig. 2a). For these dams, we applied the Google Maps geocoding API, a sophisticated cloud-based geocoding service, to retrieve the spatial coordinates of each dam as thoroughly and accurately as possible. To do so, we designed a recursive geocoding procedure that implemented three primary steps on each dam: forward geocoding, reverse geocoding, and QA filtering. The purpose of each of the steps and their logical relations are illustrated in Fig. 4.

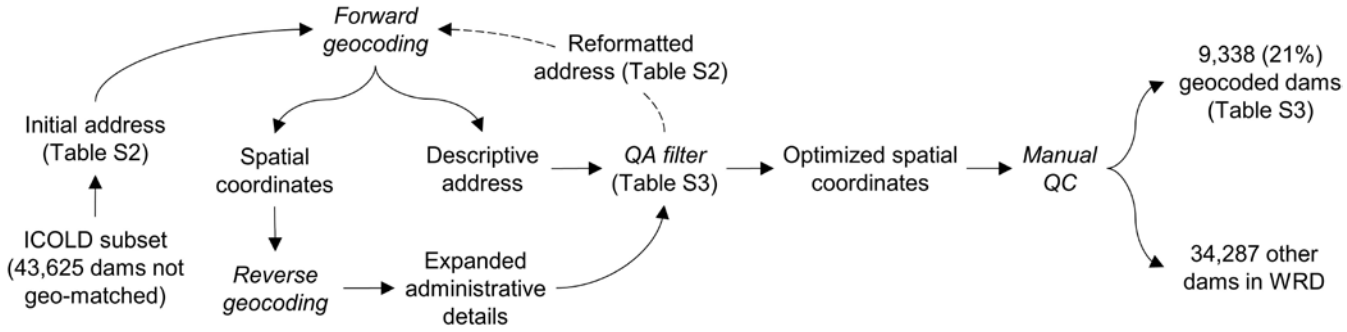


Figure 4. Schematic procedure of geocoding using Google Maps API. Text in roman indicates applied or produced datasets, and text in italics indicates methods or procedures. The dashed line arrow indicates that this step is not always necessary.

The forward geocoding (see Section 2.1 for definition) used the text address of each dam as the input, which we formatted by concatenating the WRD attribute values, to output the latitude and longitude of the dam. The WRS attributes used for address formatting include dam name, reservoir name, statement/province, and country. "Nearest town" was excluded because it is not always the township administrating the dam or reservoir. Together with the spatial coordinates, the forward geocoding also output a Google Maps address associated with the coordinates, which was parsed to individual components including feature name, street name, and political divisions. These output address components, in return, provided valuable information for QA: if the geocoded coordinates are correct, the associated output address components should agree well with those of the WRD input. However, we noticed that address components from forwarding geocoding are often limited in terms of division levels. To complement this limitation, we utilized reverse geocoding (see Section 2.2 for definition) to convert the coordinates from forward geocoding to an updated address with more complete division levels. The address components from both forward and reverse geocoding were combined and hereafter referred to as the "output address".

Similar to geo-matching, we employed a QA filter to approach the optimal geocoding result. This process first arranged the attributes of each WRD record to several address formats as they could result in different geocoding outputs. The address arrangements are listed in Supplementary Table S2, and their preference order is rationalized in Supplementary Text. Each of these WRD addresses was used iteratively for both forward and reverse geocoding (as described above). Their geocoded spatial coordinates were then ranked to five discrete QA levels based on how well the input and output addresses agree with each other (C1 to C5, Supplementary Table S3). The iteration was terminated if the highest QA rank was achieved; otherwise, the coordinates that render the best possible QA rank was used as the geocoding result.

As explained in Supplementary Table S3, the compared address components include the name of the feature and its affiliated political divisions from town/city to country levels. Consistent with geo-matching, we considered that a component was agreed on if the similarity of its values from both input and output addresses exceeds about 85%. Since the nearest town in

WRD was not used for forward geocoding, we treated it as an "independent reference" for validating the township component in the output address. Although the town or city near the dam (from WRD) does not always coincide with that administrating the dam (from the geocoding output), their occasional agreement would strengthen our confidence of the geocoded coordinates if other components were also well matched between the WRD input and the geocoding output. For this reason, we opted to include the township comparison as a supplementary criterion in the geocoding QA process. The highest QA rank (C1) corresponds to a unanimous agreement on all address components. However, the minimum rank (C5) only required the agreement on the feature name, which is a more flexible baseline in comparison with that for geo-matching. This was because some of the large reservoirs, particularly those on/near political boundaries, have shared or ambiguous divisions, and the ambiguity might be further amplified by the output coordinates which could fall in anywhere from the dam to across the reservoir water surface. In addition, some of the outputs, regardless of agreement on the address components, are not dams or reservoirs. We therefore included another baseline filter which aimed to remove any error that is not water infrastructure by analysing the feature type information in the geocoding output (see scripts in Code Availability). Although the QA process was designed to be automated, we still manually enforced hundreds of the initial outputs, many of which had returned feature names in native languages, to pass the baseline filters. As a result, our QA yielded more than 16,000 geocoded WRD records, each with the optimal spatial coordinates and the corresponding QA rank.

To complement the QA process, we then conducted a rigorous QC to correct and/or remove the remaining geocoding errors. We considered a geocoding error as a location where (a) no dam or reservoir could be visibly verified from Google Earth or Esri images, or (b) the WRD attribute information is inconsistent with the feature or division labels on Google Maps. In such cases, we usually first manually re-geocode this dam (by directly using the Google Maps interface) before deleting this error if it was not correctable. While the geo-matched coordinates from regional registers are usually on or close to the dam bodies, the geocoded coordinates could be located on the reservoir. Note the latter case was not considered as an error. Due to China's GPS shift problem, geocoded points across mainland China often show a systematic offset of roughly 500 m from their actual dam or reservoir features. For such Chinese dams, we tried to reduce their geocoding offsets by manually relocating the coordinate points to their correct dams or reservoirs. Our QC process ended up removing about 42% of the originally geocoded dams, most of which stemmed from relatively lower QA ranks (see statistics in Supplementary Table S3). The complete geocoding procedure resulted in 9,338 georeferenced and quality controlled WRD records, with an overall success rate of 21%.

2.4 Supplementation with other global inventories

The outputs from both geo-matching and geocoding, a total of 22,528 georeferenced ICOLD WRD records (Fig. 2a), was further supplemented or harmonized by two global dam/reservoir inventories to improve our inclusion of the world's largest dams. We considered this process necessary for two reasons. First, our georeferencing process, particularly geocoding via Google Maps API, did not warrant an exhaustive inclusion of the largest dams. This is particularly evident for regions where

the address and label information in Google Maps is either lacking or difficult to pass the automated QA due to language ambiguity or naming discrepancies. Second, through cross-referencing we noted that the attribute values of reservoir storage capacity provided in ICOLD WRD are occasionally erroneous (also noted by Mulligan et al. (2020)), e.g., by a factor of 1000 probably caused by unit confusion in WRD compilation. As part of the supplementation/harmonization process, we therefore collated the ICOLD reservoir storage capacities with those in the two global inventories below and corrected any evident errors in ICOLD.

2.4.1 Supplementation with Wada et al (2017): forming GeoDAR v1.0

Wada et al. (2017) compiled a list of all 144 large dams with a reservoir storage capacity larger than 10 km^3 in the world. Among them, 139 dams were provided with spatial coordinates. We verified each of the dam locations and made minor adjustments to further assure the quality. The attributes of these 139 dams were then manually compared with those in ICOLD WRD. We found that 133 of them were documented in WRD but 32 were georeferenced unsuccessfully in our geo-matching or geocoding procedure. Therefore, we borrowed the spatial coordinates of these 32 large dams in Wada et al. (2017) to supplement what we had georeferenced. The coordinates of the other 101 large dams, which we georeferenced successfully (41 from geo-matching and 60 from geocoding), were also overwritten by those in Wada et al. (2017) to double-assure and improve their spatial accuracies. This supplementation is illustrated by the Venn diagram in Fig. 2a.

We then compared the storage capacities of each of the 133 dams in Wada et al. (2017) with those in WRD and identified 22 of them exhibiting substantial discrepancies between the two datasets. We then collated their storage capacities with other documents (e.g., regional inventories, GRanD, and Wikpedia) and concluded that Wada et al. (2017) may supersede WRD in the accuracy of storage capacity for 16 of the 22 dams. The storage capacities of these 16 dams in Wada et al. (2017) were used to replace the original WRD capacities. Our data collation and verification for Wada et al. (2017) are given in Supplementary Table S4 (full spreadsheet accessible at https://doi.org/10.5281/zenodo.6163413). The entire supplementation process, including adding new dams, updating existing dam coordinates, and correcting reservoir storage capacities, increased the total storage capacity of our georeferenced dams by 15%, and 70% of the capacity increase comes from the 32 added large dams. For improved clarity, it is worth reiterating that all dams supplemented by Wada et al. (2017) were also documented in ICOLD WRD. The combined results of geo-matching and geocoding, after the supplementation from Wada et al. (2017), defines GeoDAR v1.0 containing 22,560 georeferenced records in ICOLD WRD.

2.4.2 Harmonization with GRanD: forming GeoDAR v1.1

While GeoDAR v1.0 largely exceeds GRanD in dam count, a visual comparison of their spatial distributions revealed that the latter is often complementary to (instead of completely duplicated by) the former in many regions of the world. This motivated us to perform a systematic harmonization between the two datasets. The merged version, which we entitled

GeoDAR v1.1, combines the merits of GRanD in accurately documenting the world's largest dams and GeoDAR v1.0 in providing extensive spatial details of smaller but more widespread dams.

We assumed that GRanD, by having collated multiple data sources, is superior to GeoDAR v1.0 in the accuracies of both spatial locations and attribute values (particularly reservoir storage capacity) of the world's largest dams. While this may be true for most cases, we identified at least 88 dams in GRanD with possible location errors. With the help of several references such as regional registers (Table 2), the recently published Dataset of Georeferenced Dams in South America (DDSA) (Paredes-Beltran et al., 2021), Google Maps, and other online documents, we were able to correct the locations of 76 of these dams and absorbed the corrected coordinates to the harmonization. The other 12 GRanD dams, including 3 duplicates with other dams and 9 we were unable to correct the locations for, were excluded from the harmonization. What was also excluded are another 5 dams in GranD that were subsumed or replaced by newer dams. For user convenience, we released these ~90 GranD dams together with the identified issues and suggested coordinates (if possible) in Supplementary Table S5 (full spreadsheet accessible at https://doi.org/10.5281/zenodo.6163413). Using the adjusted GRanD data (7,303 points), the harmonization (Fig. 5) aimed at (a) improving spatial coordinates of the dam points in GeoDAR v1.0, (b) adding WRD dams that are not georeferenced in GeoDAR v1.0 but are included by GRanD, (c) correcting storage capacity errors in the georeferenced WRD, and (d) absorbing the remaining GRanD dams that are not documented in WRD. Detailed processing for each of the objectives is given below.

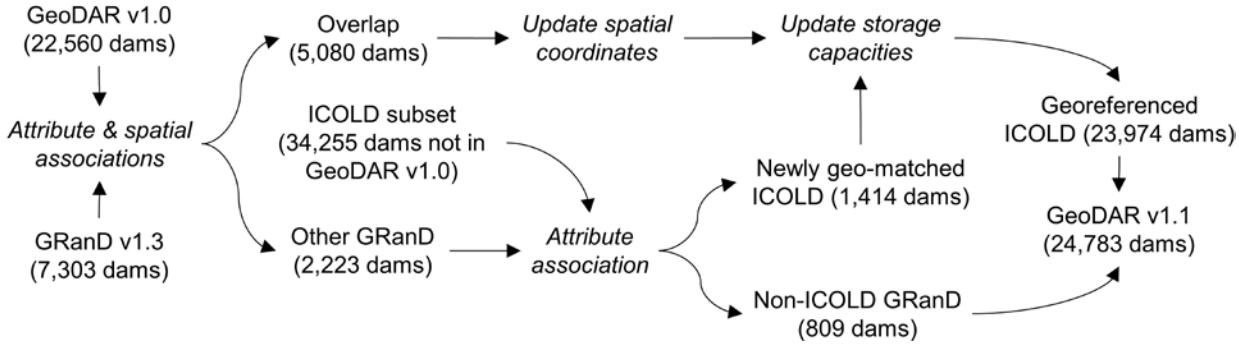


Figure 5. Schematic procedure of harmonizing GeoDAR v1.0 and GRanD v1.3 to form GeoDAR v1.1 Text in roman indicates applied or produced datasets, and text in italics indicates methods or procedures. GRanD used for harmonization excludes 74 problematic records (see Supplementary Table S5).

First, when a dam in GeoDAR v1.0 also exists in GRanD, the spatial coordinates of the former were replaced by those of the latter. We implemented a two-step procedure to identify the overlapping dams between GeoDAR v1.0 and GRanD. Step 1 was based on attribute association while Step 2 utilized spatial query. Specifically, Step 1 detected matching records between ICOLD WRD and GRanD by assessing agreements on several attributes, including dam/reservoir names, administrative

divisions, impounded rivers, and completion years. This step was essentially the same as "geo-matching" that was used to link WRD records to regional registers for GeoDAR v1.0 (Section 2.2). The association results, after a meticulous manual QC, identified ~4,670 dams in GRanD that were georeferenced in GeoDAR v1.0. For the remaining GRanD dams, Step 2 utilized their reservoir polygons to spatially intersect with the dam points in GeoDAR v1.0. A distance tolerance of ~5 km was applied to assist the spatial association and account for possible offsets in GeoDAR v1.0. As part of the QC, the attribute values of each pair (one from GRanD and the other from WRD) were manually compared to determine whether they are indeed the same dam. This step identified another 400 or so overlapping dams between the two datasets. In total, we found that GeoDAR v1.0 overlaps 5,080 out of the 7,303 dams in GRanD, and their spatial coordinates were updated to be consistent with those in GRanD.

Second, for the remaining 2,223 dams in GRanD that do not overlap GeoDAR v1.0, we assumed that at least part of them could be matched to the WRD records not georeferenced in GeoDAR v1.0. Therefore, we performed another round of attribute association between the remaining subsets of GRanD and WRD. After QC, this process identified another 1,414 WRD dams that are included by GRanD. These additional WRD dams, with a total storage capacity of 603 km^3, were then added to our inventory using the spatial coordinates from GRanD. As a result of the first two objectives, GeoDAR v1.1 georeferenced 23,974 (42%) out of the 56,815 dams in ICOLD WRD, including6,494 that overlap with GRanD.

Third, to reduce the impact of possible attribute errors in ICOLD WRD, we next merged the values of reservoir storage capacity from both WRD and GRanD to a single updated attribute, where the original values in WRD or Wada et al. (2017) were overwritten by those of the overlapping dams in GRanD (if the GRanD values are valid). This correction led to a minor increase of 86 km^3 (1.2%) in the total reservoir storage capacity. Eventually, the remaining 809 dams in GRanD, which were not found in WRD, were appended to our georeferenced WRD so that the final inventory absorbed the entire dataset of GRanD. It is worth noting that similar to geo-matching (Section 2.2), our attribute association here could be conservative, meaning that some of the dams appended from GRanD might be documented in the remaining WRD (the subset not georeferenced successfully). The complete harmonization process, combining the above three steps, led to a total of 24,783 georeferenced dams in GeoDAR v1.1 (Fig. 2b).

2.5 Retrieving reservoir boundaries

Reservoir polygons of the georeferenced dam points were retrieved as thoroughly as possible from three global water body datasets: GRanD reservoirs (Lehner et al., 2011), HydroLAKES v1.0 (Messager et al., 2016), and UCLA Circa 2015 Lake Inventory (Sheng et al., 2016). These three water body datasets exhibit an increasing spatial resolution: from 7000+ polygons in GRanD reservoirs provided exclusively for GRanD's dam points, to millions of water body polygons, including both natural lakes and reservoirs, in the other two datasets. While HydroLAKES documents 1.4 million water bodies larger than 0.1 km^2 (10 ha), the Landsat-based UCLA Circa 2015 Lake Inventory further reduced the minimum size to only 0.004 km^2

(0.4 ha), resulting in another 7.7 million water bodies on the global continental surface. Accordingly, we implemented a hierarchical procedure, where the three water body datasets were applied in ascending order of spatial resolution to retrieve the reservoir boundaries with an overall decreasing size.

Specifically, GRanD v1.3 provides 7,230 reservoir polygons for the 7,303 dam points used for harmonization. These GRanD polygons were first assigned to their associated dam points in GeoDAR v1.1 through GRanD IDs. Reservoirs of the remaining 17,480 dam points in GeoDAR v1.1, which were georeferenced from ICOLD alone, were next retrieved from HydroLAKES when possible. To avoid duplicates in the reservoirs retrieved from different data sources, we only used the subset of HydroLAKES that is spatially independent from (i.e., not intersecting with) GRanD reservoirs. Different from reservoir assignment using GRanD, there was no common attribute ID to pair HydroLAKES polygons with the remaining dam points, so their reservoir retrieval relied completely on spatial association. One major challenge in dam-reservoir spatial association was the ambiguity caused by the offsets between our georeferenced dam points and their actual reservoir polygons (see Section 2.3).

To tackle this ambiguity, we designed a procedure that consists of three rounds of iteration to progressively optimize reservoir-dam association. This procedure was based on two assumptions, both conditional on a reasonable spatial tolerance. We started with 500 m to be roughly consistent with the georeferencing offset observed in China. The first assumption was that larger reservoirs are more likely to be documented than smaller ones, in both ICOLD WRD and Google Maps. Therefore, the first round of iteration assigned each of the dams to the largest water body within the tolerance. This assignment might, however, lead to a situation where multiple dams were assigned to the same reservoir. To untangle this situation, the remaining iterations assumed Tobler's First Law of Geography (Tobler, 1970): "everything is related to everything else, but near things are more related than distant things" (p.236). Accordingly, for any water body mistakenly associated with multiple dams, the second round of iteration reassigned the water body to its closest dam, and the other dam(s) within the tolerance, as a result, was/were left unpaired. To reduce the number of such "orphan" dams, a final, third round of iteration assigned the remaining unpaired dams to the next closest water body that was within the spatial tolerance and had not been previously associated with any dams. If this led to multiple dams associated with one reservoir again, only the dam with the closest proximity to the reservoir was kept. Through experimentation, we opted to implement this three-iteration procedure twice, first using a conservative 500-m tolerance to maximize the accuracy for most associations, and then a 1-km tolerance to further minimize the number of orphan dams.

This multi-iteration procedure retrieved roughly 7,600 reservoir polygons from HydroLAKES. For the remaining dam points left unpaired, we applied the same association procedure to continue retrieving their reservoirs from the high-resolution UCLA Circa 2015 Lake Inventory. Similarly, only the subset that does not intersect with the retrieved HydroLAKES

polygons was considered, in order to avoid duplicates in the retrieved reservoirs from different datasets. The use of UCLA Circa 2015 Lake Inventory retrieved another 6700 or so reservoirs.

We followed the automated reservoir retrieval by a manual QC to visually 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. This visual QC was particularly necessary for lake-dense regions, including the case of cascade reservoirs immediately downstream/upstream to each other. While some of the dams, such as barrages, diversion infrastructure, and dams under construction, do not have visible impoundments (Lehner et al., 2011), we tried to be as meticulous as possible to verify and recover any missing reservoirs. For instance, we were unable to manually retrieve 10 reservoirs (including 4 completed after 2000) from the UCLA Circa 2015 Lake Inventory for the ~70 dams in GRanD v1.3 without polygons. We also replaced hundreds of reservoirs initially retrived from GRanD and HydroLAKES by the polygons in the UCLA inventory to improve the boundary accuracy and completeness.

3 Product components and usage

We here provide a detailed documentation of the components and structure of the GeoDAR versions (v1.0 and v1.1). To facilitate the description, the two GeoDAR versions and their component statistics are explained in Table 1, and spatial distributions of the dam points and reservoir polygons are visualized in Figs. 6 and 7.

3.1 GeoDAR v1.0: dams

GeoDAR v1.0 is a collection of 22,560 dam points georeferenced exclusively for ICOLD WRD (Fig. 6a). Among them, 13,149 or 58% were retrieved from geo-matching regional dam registers, 9,278 or 41% from Google Maps geocoding API, and the remaining 133 largest dams from the spatial inventory in Wada et al. (2017) (Fig. 6b). WRD storage capacities of most of these 133 large reservoirs were replaced by the values in Wada et al. (2017) (see Section 2.4.1), and unless stated otherwise, our following statistics on storage capacities were calculated after this replacement.

The total reservoir storage capacity of these dams is 6,441 km^3, meaning that GeoDAR v1.0 georeferenced 40% of the 56,815 WRD records but included more than 80% of their cumulative reservoir storage capacity. The total storage capacity of the 133 largest dams from Wada et al. (2017), despite being limited in number, reaches 3900 km^3 or 61% of the cumulative storage capacity in GeoDAR v1.0, and the other ~40% capacity was split almost equally between the remaining 22,000+ geo-matched and geocoded dams. Although the registers used for geo-matching are regional, the dams in GeoDAR v1.0, as shown in Fig. 6b, are distributed in 151 out of the 165 countries or territories in WRD, largely owing to our geocoding efforts through Google Maps API. Since GeoDAR v1.0 was produced independently from other global dam

datasets such as GRanD, it can also be used to cross-compare, supplement, and potentially improve other dam datasets. Validation of our georeferencing accuracy for v1.0 is provided in Section 4.

3.2 GeoDAR v1.1: dams and reservoirs

GeoDAR v1.1 consists of a) 24,783 dam points (Fig. 6a) representing a full harmonization between GeoDAR v1.0 and GRanD v1.3, and b) 21,515 reservoir polygons (Fig. 7). In these nearly 25,000 dam points, 17,480 or 71% come from GeoDAR v1.0 alone, 6,494 or 26% shared by ICOLD WRD and GRanD, and the other 809 or 3% from GRanD alone (Table 1; Fig. 6c). Among the 6,494 shared dams, 5,080 were georeferenced in both GeoDAR v1.0 and GRanD, and the remaining 1,414 were introduced through the harmonization with GRanD. This resulted in a total of 23,974 georeferenced WRD records (42% of all WRD records) in GeoDAR v1.1. In addition to the expanded number of georeferenced WRD dams, GRanD supplemented another 809 dams which are exclusive of WRD. The total 2,223 dams added by GRanD, notated as "GRanD v1.3 & other ICOLD" and "GRanD v1.3 only" in Fig. 6c, are distributed worldwide and complement v1.0, particularly in regions such as Africa and Central Asia where geocoding using Google Maps was challenging. After this ICOLD-GRanD harmonization, the spatial coverage of the dam points in GeoDAR v1.1 increased to 155 out of the 165 countries in WRD.

As described in Section 2.4.2, we substituted the reservoir storage capacities in GRanD for the original capacity values of their overlapping WRD dams. As a result, the total reservoir storage capacity in GeoDAR v1.1 reaches 7,384 km^3, which compares to ~95% of the cumulative capacity in the entire ICOLD WRD (see Section 5.1 for more comparisons with ICOLD). As reported in Table 1, 81% (6,006 km^3) of the total storage capacity in GeoDAR v1.1 is explained by the 5,080 relatively large dams georeferenced in both GeoDAR v1.0 and GRanD. The 17,480 smaller dams from GeoDAR v1.0 alone contribute only 7% (507 km^3) of the total storage capacity, which is roughly comparable to the subset from GRanD alone (268 km^3) or the subset from GRanD and other ICOLD WRD (603 km^3). These capacity contributions suggest that compared to GRanD, the major improvement of GeoDAR lies on the increased number of relatively small dams, rather than the increase in total storage capacity of the dams (see Section 5.2 for more comparisons with GRanD).

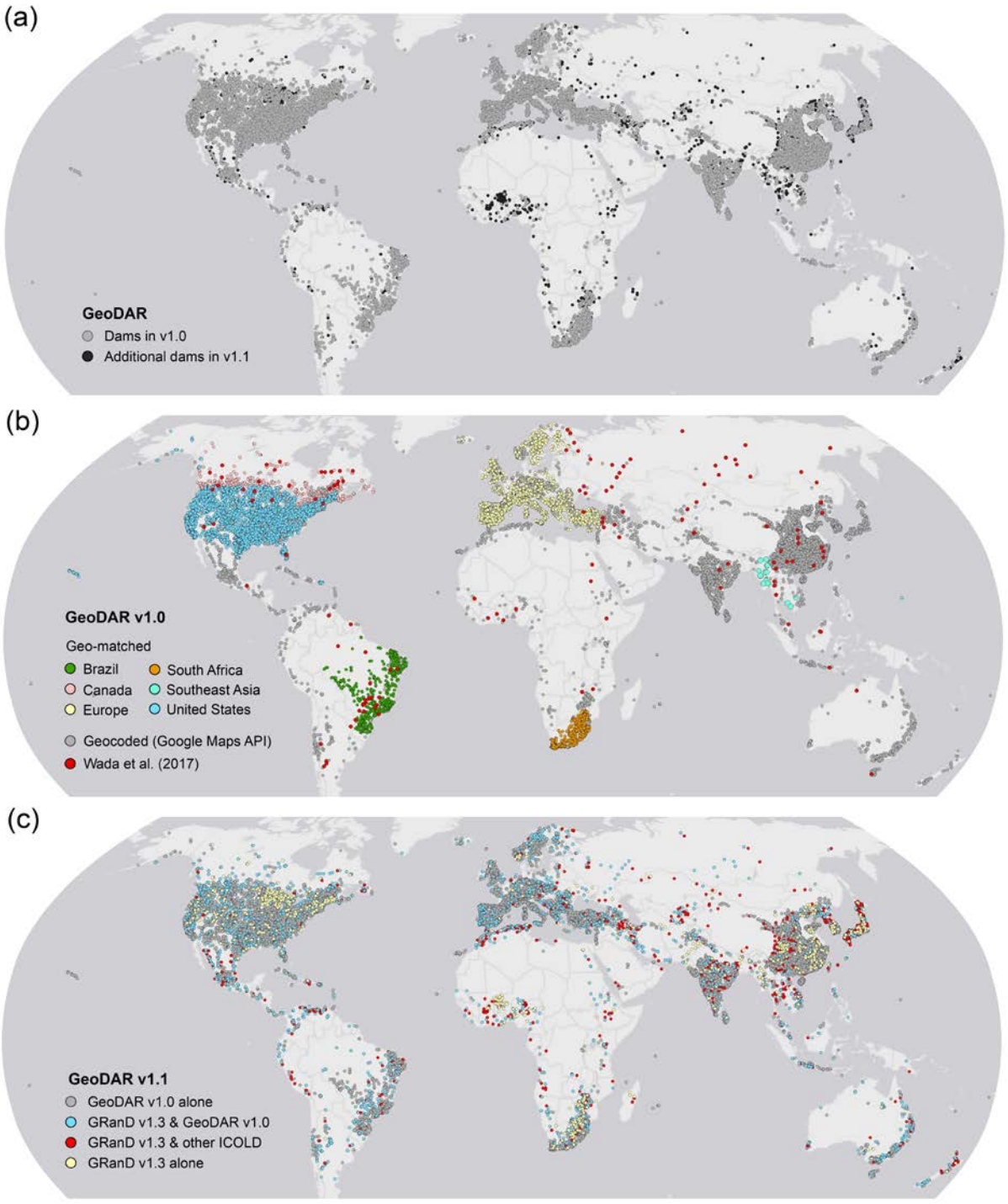


Figure 6. Georeferenced dam points in GeoDAR. (a) A total of 24,783 dam points in v1.1 superimposed by 22,560 dam points by in v1.0. (b) Georeferencing methods and data sources for v1.0. (c) Data sources for v1.1.

Different from GeoDAR v1.0, version 1.1 also includes a component of reservoir polygons which represent water impoundment extents associated with 21,515 or 87% of the georeferenced dam points (Fig. 7). Reservoir polygons for the remaining 13% of the dam points were retrieved unsuccessfully due to a combination of factors, including limited spatial resolutions of the applied water masks, offsets in our georeferenced dam points, and the fact that some of the dams have no evident water impoundments. Nevertheless, the retrieved reservoir polygons have a cumulative area of 496,314 km^2, accounting for 98% of the total reservoir area of all georeferenced dams in GeoDAR v1.1 (reservoir areas without polygons are based on documented attributes). These retrieved reservoirs correspond to a cumulative storage capacity of 7,216 km^3, also accounting for nearly 98% of the total storage capacity in v1.1. These statistics indicate that the reservoirs whose boundaries were retrieved unsuccessfully were mostly small in area and storage.

The numbers of reservoir polygons retrieved from each of the three water body datasets are fairly comparable (about 7,100–7,200 each), but the total reservoir storage capacity and area decrease with the increasing spatial resolution of the water body datasets (Table 1). As a result, the mean reservoir polygon size decreased from 63 km^2 for those retrieved from GRanD, to 2 km^2 from HydroLAKES and 5 km^2 from the UCLA Circa 2015 Lake Inventory. This result is overall consistent with the design of our hierarchical procedure (Section 2.5), where smaller reservoirs were successively retrieved with the help of finer water masks. It is important to note that the retrieved polygons do not always represent the maximum water extents of the reservoirs because water boundaries in the retrieval sources were not necessarily mapped in the maximum inundation periods. For example, the UCLA Circa 2015 Lake Inventory contains more than 9 million water bodies larger than 0.4 ha, which were mapped from Landsat images acquired during the “steady” climate periods (Lyons and Sheng, 2018) and thus represent the average seasonal extent of each water body (Sheng et al., 2016). Despite not always being the largest water extents, our retrieved reservoir polygons enhanced the spatial details of global reservoir locations, using which users can further expand or refine the water boundaries to their specific needs.

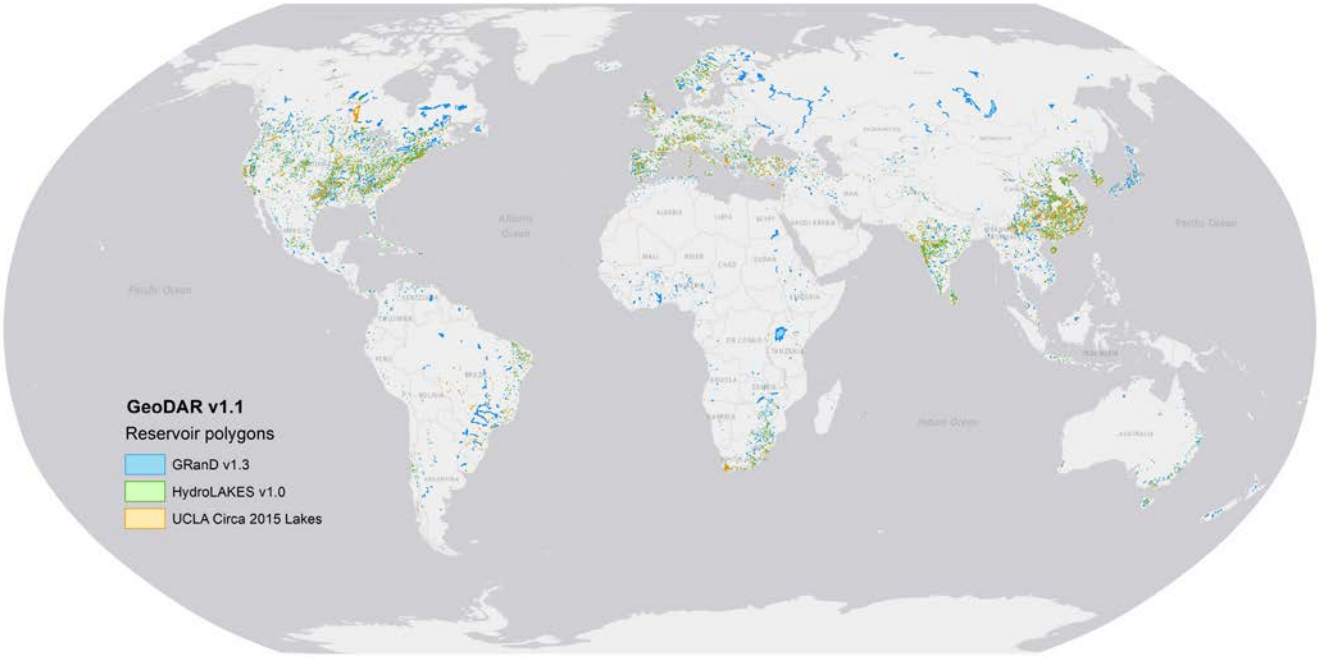


Figure 7. Reservoir polygons and their retrieval data sources in GeoDAR v1.1. For display, GRanD polygons are superimposed by HydroLAKES polygons and then by UCLA Circa 2015 Lakes.

3.3 Attributes and usage

The GeoDAR dataset, including dam points for v1.0 and both dam points and reservoir polygons for v1.1, is provided as three separate shapefiles. For user convenience, we also duplicated the two dam point shapefiles in the comma-separated values (csv) format. The file names and attributes are explained in Table 3. Although most of our dam points were georeferenced using WRD records, our published GeoDAR complies with the proprietary rights of ICOLD and does not directly release any attribute from WRD. The attributes we provide in GeoDAR, as listed in Table 3, are only limited to our georeferencing methods, QA/QC, validation, and other information (such as spatial coordinates and part of the reservoir storage capacities) that is already open source or has been permitted for use by the original producers.

Table 3. Attributes in the data products of GeoDAR

Attribute	Description and values
v1.0 dams (file name: GeoDAR_v10_dams; format: comma-separated values (csv) and point shapefile)	
id_v10	Dam ID of GeoDAR version 1.0 (type: integer). Note this is not the "International Code" in ICOLD WRD but is associated with "International Code" through encryption.
lat	Latitude of the dam point in decimal degree (type: float) on datum World Geodetic System (WGS) 1984.
lon	Longitude of the dam point in decimal degree (type: float) on WGS 1984.

geo_mtd	Georeferencing methods (type: text). Unique values include: “geo-matching CanVec”, “geo-matching LRD”, “geo-matching MARS”, “geo-matching NID”, “geo-matching ODC”, “geo-matching ODM”, “geo-matching RSB”, “geocoding (Google Maps)”, and “Wada et al. (2017)”. Refer to Table 2 for acronyms.
qa_rank	Quality assurance (QA) levels (type: text). Unique values include: “M1”, “M2”, “M3”, “C1”, “C2”, “C3”, “C4”, and “C5”. Refer to Supplementary Tables S1 and S3 for explanation.
rv_mcm	Reservoir storage capacity or volume in million cubic meters (type: float). Values are only available for dams acquired from Wada et al. (2017). Capacity values of other records in ICOLD WRD are not released due to proprietary restriction.
val_scn	Validation result (type: text). Unique values include: “correct”, “register”, “mismatch”, “misplacement”, and “Google Maps”. Refer to Table 4 for value explanation.
val_src	Main sources used for validation (type: text). Values include: “CanVec”, “Google Maps”, “JDF”, “LRD”, “MARS”, “NID”, “NPCGIS”, “NRLD”, “ODC”, “ODM”, “RSB”, and “Wada et al. (2017)”. Refer to Table 2 for acronyms.
qc	Roles and name initials of co-authors/personnel participating in data quality control (QC) and validation.
v1.1 dams (file name: GeoDAR_v11_dams; format: comma-separated values (csv) and point shapefile)	
id_v11	Dam ID of GeoDAR version 1.1 (type: integer). Note this is not the “International Code” in ICOLD WRD but is associated with “International Code” through encryption.
id_v10	v1.0 ID of this dam/reservoir (as in *ID_v10*) if it is also included in v1.0 (type: integer).
id_grd_v13	GRanD ID of this dam if also included in GRanD v1.3 (type: integer).
lat	Latitude of the dam point in decimal degree (type: float) on WGS 1984. Value may be different from that in v1.0.
lon	Longitude of the dam point in decimal degree (type: float) on WGS 1984. Value may be different from that in v1.0.
geo_mtd	Same as *geomtd* in v1.0 if this dam is included in v1.0.
qa_rank	Same as *QA_level* in v1.0 if this dam is included in v1.0.
val_scn	Same as *val_scn* in v1.0 if this dam is included in v1.0.
val_src	Same as *val_src* in v1.0 if this dam is included in v1.0.
rv_mcm_v10	Same as *rv_mcm* in v1.0 if this dam is included in v1.0.
rv_mcm_v11	Reservoir storage capacity in million cubic meters in this version (type: float). Due to proprietary restriction, provided values are limited to dams in Wada et al. (2017) and GRanD v1.3. If a dam is included by both Wada et al. (2017) and GRanD v1.3, the value from the latter takes precedence.
har_src	Source(s) to harmonize the dam points. Unique values include: “GeoDAR v1.0 alone”, “GRanD v1.3 and GeoDAR 1.0”, “GRanD v1.3 and other ICOLD”, “GRanD v1.3 alone”. Refer to Table 1 for more details.
pnt_src	Source(s) of the dam point spatial coordinates. Unique values include: “GeoDAR v1.0”, “original GRanD”, “adjusted GRanD” (meaning original points in GRanD adjusted to improve the locations), “corrected GRanD” (meaning misplaced dam points in GRanD corrected and relocated; also see Table S5).
qc	Roles and name initials of co-authors/personnel for data QC, validation, and other manual operations.
v1.1 reservoirs (file name: GeoDAR_v11_reservoirs; format: polygon shapefile)	
plg_src	Source of the retrieved reservoir polygon (type: text). Unique values include “GRanD v1.3 reservoirs”, “HydroLAKES v1.0”, and “UCLA Circa 2015 Lakes”. Refer to Table 1 for more details.
plg_a_km2	Area of the retrieved reservoir polygon in square kilometres (calculated using the cylindrical equal area projection on WGS 1984).
All other attributes in v1.1 dams.	

Note: Missing or inapplicable values are flagged by “Null” for text-type attributes and “-999” for numeric-type attributes.

Although WRD attributes are not directly available in GeoDAR, we suggest two possible ways for users to acquire at least some of the essential attributes. Upon the user's reasonable request and on a case-by-case basis, we may provide assistance in decrypting the association between GeoDAR IDs (Table 3) and ICOLD's International Codes, and using the International Codes, the user can link each of the dams/reservoirs in GeoDAR to the entire 40 or so proprietary attributes in WRD. This is also based on the premise that the user needs to acquire the WRD attribute data from ICOLD, and that the user agrees not to release the GeoDAR-WRD association or the WRD attributes to the public. Alternatively, since we imposed no usage restrictions on our spatial features (geometric dam points and reservoir polygons), users are free to integrate them with other datasets and tools, such as remote sensing observations and modelling, to acquire the needed attributes, particularly those not yet documented in ICOLD WRD. Acquisition methods have been exemplified for at least the following attributes: reservoir hypsometry and bathymetry (Li et al., 2020; Yigzaw et al., 2018), surface evaporation loss (Mady et al., 2020; Zhan et al., 2019; Zhao and Gao, 2019a), operation rules (Shin et al., 2019; Yassin et al., 2019), completion years (Zhang et al., 2019), storage capacities (Liu et al., 2020), and the changes in water area (Pekel et al., 2016; Yao et al., 2019; Zhao and Gao, 2019b), level (Cretaux et al., 2011; Schwatke et al., 2015), and storage or volume (Busker et al., 2019; Cretaux et al., 2016; Gao et al., 2012; Zhang et al., 2014).

4 Validation

Separate from the QA/QC during data production, we performed a posterior validation to further assess the accuracy of the georeferenced ICOLD WRD records. The validation sample consists of about 1400 dam points (Fig. 8), which were selected worldwide from GeoDAR v1.0 and represent the results of our geo-matching and geocoding prior to GRanD harmonization. The collection of the validation points followed a stratified sampling method (Table 4). From the subset of GeoDAR v1.0 produced by geo-matching, we randomly selected about 40 dam points for each of the geo-matching regions (Brazil, Canada, Europe, South Africa, and United States), with the exception of Southeast Asia (Cambodia and Laos) where all 17 geo-matched WRD dams were included for validation. We allowed the sample to occasionally overlap with GRanD because dams in GeoDAR v1.0 were georeferenced independently from GRanD and those shared with GRanD reflect our georeferencing accuracy for the world's largest dams. However, for each regional sample, we limited the number of GRanD-overlapping dams to no more than 30% of the entire regional sample size if possible (Table 4). This was to comply with the size ratio between GRanD and GeoDAR v1.0 (about 1:3) so that our validation still emphasized smaller, newly georeferenced dams. We also randomly selected 40 out of the 133 large WRD dams supplemented by Wada et al. (2017), considering that they are part of GeoDAR v1.0 and the supplementation was based on attribute association similar to regional geo-matching. In total, 260 dams were selected for validating the geo-matching accuracy. For each dam, we manually checked whether its spatial coordinates in GeoDAR v1.0 are consistent with those documented in the geo-matching source (see source references in Table 2).

Table 4. Validation statistics for GeoDAR v1.0

Region	Main reference	Sample	Accuracy	Error source
Geo-matching		**260; 84**	**252 (96.9%)**	---
Brazil	RSB	40; 7	38 (95.0%)	Register
Canada	CanVec	41; 13	38 (92.7%)	Register; Mismatch
Europe	MARS	41; 3	40 (97.6%)	Register
South Africa	LRD	40; 11	40 (100%)	---
Southeast Asia	ODC; ODM	17 (all); 4	15 (88.2%)	Register
United States	NID	41; 9	40 (100%)	---
Global	Wada et al (2017)	40; 37	40 (100%)	---
Geocoding		**1,152; 316**	**1,094 (95.0%)**	---
China	NPCGIS	250; 30	247 (98.8%)	Misplacement
India	NRLD	220; 57	215 (97.7%)	Misplacement
Japan	JDF	232 (all); 148	209 (90.1%)	Misplacement; Google Maps
Others	Google Maps	450; 81	423 (94.0%)	Misplacement
Total		**1,412; 400**	**1,346 (95.3%)**	---

Note: In "Sample", the two numbers delimited by semicolon indicate the size of the validation sample from GeoDAR v1.0 (left) and the number of dams in this sample that overlap with GRanD v1.3 (right), respectively. "Error source" lists error scenarios in decreasing order of frequency. "Mismatch" indicates geo-matching errors due to incorrect association between WRD and the source/reference register. "Register" indicates geo-matching errors due to inaccurate spatial coordinates in the source register (despite correct association). "Misplacement" indicates geocoding errors where the WRD attribute information disagrees with the Google Maps label. "Google Maps" indicates geocoding errors due to endogenous feature labelling mistakes in Google Maps (despite the WRD attribute information and the Goole Maps label agreeing with each other). See Table 2 (column "Register/Source") for reference details.

From the remaining subset of GeoDAR v1.0 produced by geocoding, we followed the same stratified sampling scheme and selected 220 to 250 dam points for each of China, India, and Japan. Another 450 dam points were sampled from the other regions of the world (Table 4). Compared to geo-matching which was based on attribute association with georeferenced regional registers, the geocoding process was more complicated and relied largely on the geographic information repository in Google Maps and its embedded geocoding algorithms. To increase our confidence in the geocoding results, we therefore purposefully enlarged the sample size for each validation region. As described in Section 2.2, three additional georeferenced datasets from authoritative registries in China, Indian, and Japan were used exclusively for the purpose of geocoding validation (refer to Table 2 for register details). For the remaining regions of the world, the validation was based on a meticulous manual comparison between the WRD information of each sampled dam point and the associated Google Maps label, including the dam/reservoir name, administrative divisions, the nearest town/city, and the impounded river name if possible. When necessary, we also referred to other auxiliary information including open-source gazetteers and other

literature. In total, we collected 1,152 dam points for validating the accuracy of geocoding, including all 232 Japanese dams in GeoDAR v1.0. The distribution of all sampled validation dams is shown in Fig. 8.

As reported in Table 4, our geo-matching accuracy ranges from 88% to 100% among different regions, with an overall accuracy of 97%. Causes of the identified geo-matching errors (see the last column in Table 4) were not necessarily mistakes in our attribute association between WRD and the georeferenced registers, but sometimes inaccurate spatial coordinates provided by the georeferenced registers themselves. An example is Skutvik Dam (completion year 1991) in Norway (Fig. 8), where coordinates are documented to be 68.025° N and 15.345° E in MARS. However, inspected from high-resolution Google Maps imagery, no dam or reservoir could be conclusively verified at or near this coordinate point, except for three surrounding lakes that are all over 2 km away and labelled with other names (Vanbassenget, Lanstøvatnet, and Stenslandsvatnet). The documented coordinates for this dam are probably inaccurate.

The accuracies of our geocoded samples ranges from 90% for Japan to 98–99% for India and China, with an overall accuracy of 95%. As shown in Table 4, most of the errors were related to the misplacement of the dam/reservoir to another feature, typically a free-flowing river reach, which shares the name and administrative divisions with the dam/reservoir. One example is Nambiar Dam near the city of Tirunelveli in the state of Tamil Nadu, southern India (Fig. 8). The correct coordinates, according to NRLD, are 8.374° N and 77.738° E where the Google Maps labelled “Nambi Dam” instead of Nambiar Dam. Probably because of this spelling inconsistency, our geocoded coordinates were misplaced on a reach of the Nambi(y)ar River (8.435° N, 77.569° E, labelled as “Nambiyar”) about 20 km upstream from the dam. Although our recursive geocoding procedure (Section 2.3) embedded an automated filter that examines the type of the feature at each returned point, this filter was designed to only eliminate the coordinates where feature types are clearly disparate from a dam or reservoir (such as commercial and residential buildings). Our experiments showed that dams/reservoirs and free-flowing river reaches could both be categorized as “establishment” of “natural feature” and a feature type that is more specific to dams/reservoirs was hardly seen. Thus, to avoid over-filtering, we allowed a certain ambiguity in the geocoded feature types, and then relied on manual QC to correct or remove mistaken coordinates as thoroughly as possible. The misplacement of dams to their upstream/downstream river reaches is a major cause of the relatively low geocoding accuracy in Japan. Through experimentations, we noticed that Google Maps labelling for some of the Japanese dams that are homonymous to their impounded rivers, is either lacking or highly adapted to the Japanese language. The latter further challenged our geocoding accuracy using English-based ICOLD information. For one of the errors in Japan, we verified from the JDF register that Google Maps mislabelled Myojin Dam in Horoshima Prefecture (34.587° N, 132.505° E) as “Nabara Dam” whose correct location is 3 km downstream (34.563° N, 132.517° E; Fig. 8). As a result, our georeferenced coordinates for Nabara Dam were wrong although our geocoding process was correct. However, given what we have observed, such endogenous labelling errors in Google Maps are probably rare.

Integrating the validations for both geo-matching and geocoding, our overall georeferencing accuracy is 95.3% in terms of dam count or 99.1% in terms of total storage capacity based on the sampled 1,412 dams. While these statistics can be considered as an accuracy measure of our data product, the identified errors in the validation sample have been corrected wherever possible or otherwise removed in our released GeoDAR v1.0 and v1.1 (for simplicity, our reported statistics after QC have considered this additional correction). To reflect the accuracy of GRanD harmonization, we also randomly sampled another ~100 dams in v1.0 that were associated with GRanD in v1.1, and identified no association errors among them.

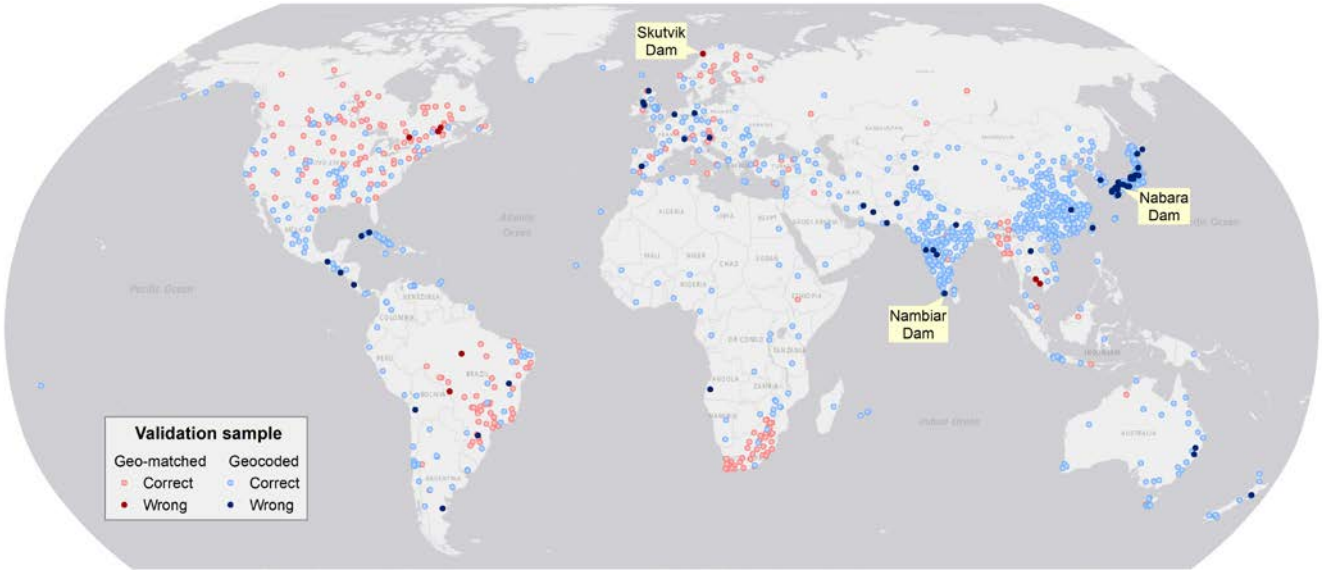


Figure 8. Validation sample and results for GeoDAR v1.0. The validation sample consists of 1,412 georeferenced ICOLD dams, including 260 dams from geo-matching and 1,152 dams from geocoding. See Table 4 for detailed validation statistics.

5 Comparisons with existing global datasets

To better understand the improvements and potential applications of GeoDAR, we compare it with three major global dam and reservoir datasets: the complete ICOLD WRD, GRanD (v1.3), and GOODD (V1). To recap the pros and cons of each dataset, ICOLD WRD documents over 56,000 unique dam records with a broad suite of attributes, but the provided dam records are not georeferenced. GOODD depicts the spatial details of more than 38,000 dam points and their catchments but does not include any other attribute. GRanD is georeferenced and provides multiple essential attributes, but the records are limited to 7320 large dams. Accordingly, our comparison first emphasized the aspects of dam quantity, reservoir area, and if applicable, the spatial pattern and distribution of the dams. These aspects are openly available from the spatial features (i.e., dam points and reservoir polygons) in GeoDAR. Considering that each GeoDAR feature is also linked to a WRD or GRanD

620 record which contains detailed attributes, our comparison also includes two important attributes, i.e., reservoir storage capacity and catchment area, to help inform the extended capability of GeoDAR once the attributes are acquired.

5.1 Comparison with ICOLD WRD

Despite our efforts to integrate multi-source registers and the Google Maps geocoding API, georeferencing ICOLD WRD, particularly smaller dams in poorly documented regions, has proven to be challenging. This challenge was reflected by the
625 proportion of WRD that was spatially resolved in GeoDAR. As compared in Table 5, GeoDAR v1.0 included 40% of the 56,815 records in the entire WRD. Although limited in number, these georeferenced records compromised a balance between geocoding thoroughness and quality (see Sections 2.2 and 2.3), and account for 84% of the total reservoir storage capacity in WRD. The larger proportion in terms of storage capacity indicates that most of the sizable dams in WRD have been spatially resolved. This message is also corroborated by Fig. 9. Nearly 70% of the 12,412 WRD dams larger than 10
630 mcm, for example, have been georeferenced in GeoDAR v1.0 (Fig. 9a). While 80% of the 21,849 WRD dams smaller than 1 mcm were not georeferenced, these smaller dams account for less than 1% of the total WRD storage capacity (Fig. 9b). After harmonization with GRanD, the proportion of WRD georeferenced in GeoDAR v1.1 increased to 42% by count or 92% by storage capacity (Table 5), and these percentages represent our best result for georeferencing WRD. By absorbing the remaining dams in GRanD as well, v1.1 has a total dam count equivalent to 44% of WRD and a cumulative storage capacity
635 less than 5% below that of the full WRD (Table 5; Fig. 9b). Compared to v1.0, the margin between the distribution curves of GeoDAR v1.1 and WRD, particularly for relatively large dams, was further reduced (Fig. 9a). As a result, the number of dams larger than 10 mcm in GeoDAR v1.1 exceeds 80% of that in WRD, and the number of dams larger than 1 mcm reaches 60% of that in WRD.

Table 5. Summative comparisons among GeoDAR, ICOLD, and GRanD

Statistics	ICOLD	GRanD	GeoDAR		
	Full WRD	v1.3	v1.0 (WRD)	v1.1 (WRD)	v1.1 (WRD ∪ GRanD)
Dam count	56,815	7,320	22,560	23,974	24,783
Storage capacity (km^3)	7,720.2	6,881.0	6513.2	7,116.2	7,383.8
Reservoir area (km^2)	519,159.5	475,543.9	---	476,602.5	496,313.8
Catchment area (10^3 km^2)	150,114.6	116,455.9	---	140,389.4	147,958.1

640 Note: To improve the validity of data comparison, statistics in Seciton 5 are based on the following adjustments. When a dam is documented in both GRanD and WRD, attribute values in GRanD (if available) took precedence (meaning that WRD values were replaced by those in GRanD). Exceptions are the GRanD dams not used in GeoDAR v1.1 harmonization (Supplementary Table S5). When a dam has both a reservoir polygon and an area attribute, the polygon area took precedence

for calculating “Reservoir area” statistics. Reservoir area statistics for GeoDAR v1.1 only considered the dams whose reservoir polygons were successfully retrieved. Statistics for GRanD are based on the entire records in v1.3.

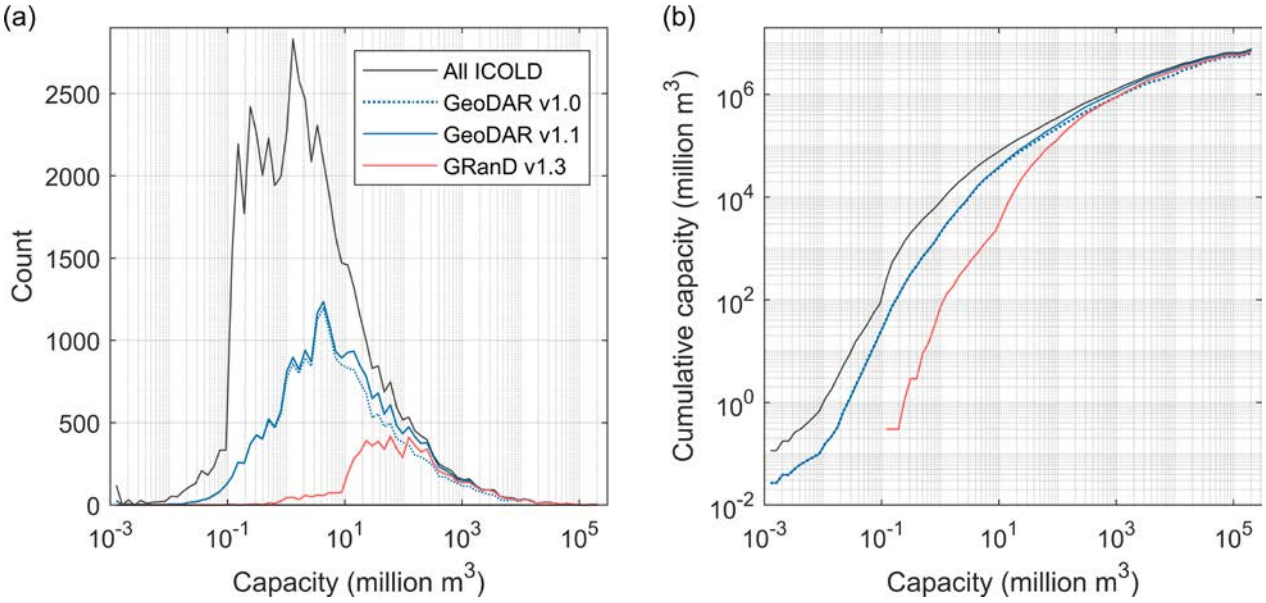


Figure 9. Comparison among GeoDAR, ICOLD WRD, and GRanD by reservoir storage capacity. (a) Frequency (count) distribution. (b) Cumulative (integral) storage capacities. Statistics were based on 80 equal-size bins on a logarithmic scale between the minimum and maximum storage capacities (i.e., 0.001 to 204,800 mcm).

The spatial coverage of GeoDAR, in comparison with WRD, was summarized for each of the 165 countries with registered WRD records (Fig. 10). Our comparison focused on GeoDAR v1.1 as it represents an improved version of our spatial dam inventory. Among these 165 countries, the median proportion of the dam count covered by GeoDAR v1.1 is 62%, with the first and third quartiles being 35% and 89%, respectively. As shown in Fig. 10a, better coverages tend to occur in North America, Europe, Russia, Australia, and part of South America and Africa, whereas poorer coverages are seen in East Asia, South Asia, and part of the Middle East. The coverages in China and India, for example, are only about 22–26% due to a large quantity of WRD records for these two countries (23,749 in China excluding Taiwan, and 5,074 in India) but relatively limited information on Google Maps. Compared with dam counts, GeoDAR’s coverage for reservoir storage capacity is higher overall (Fig. 10b). Among the 157 countries with documented reservoir storage capacities, the median coverage in GeoDAR reaches 98%, with the first and third quartiles being 87% and 100%, respectively. If we exclude the 809 dams supplemented by GRanD alone and only consider the WRD portion of GeoDAR v1.1, the coverage becomes overall lower but by a limited extent. Among these countries, the median proportion of the WRD dams covered by the WRD portion of GeoDAR v1.1 is 59% (with 33% and 83% as the first and third quartiles) in terms of dam count and 95% (82% and over 99% as first and third quartiles) in terms of reservoir storage capacity (Supplementary Fig. S1), suggesting that a substantial

proportion of WRD had been georeferenced in many of the register countries before the additional supplementation from GRanD. More detailed comparisons (among ICOLD, GranD v1.3, and GeoDAR v1.3) for each of the 165 countries are given in Supplementary Table S6.

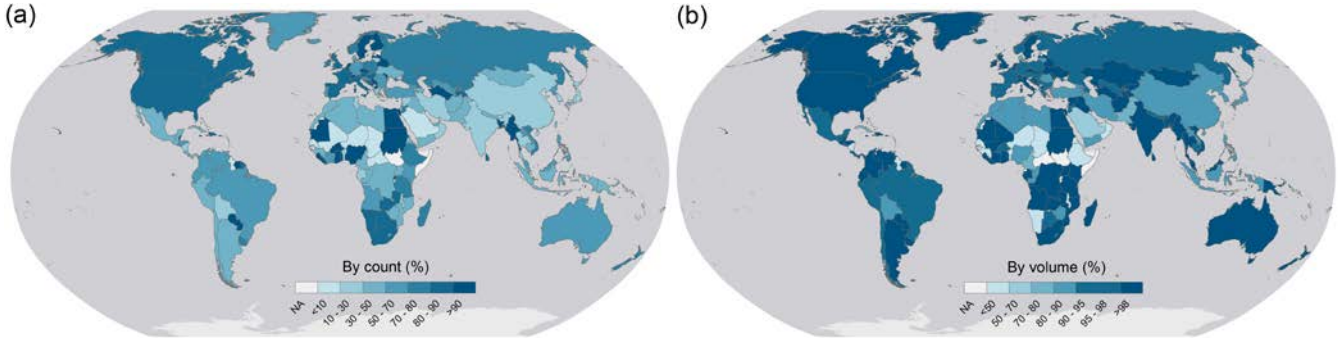


Figure 10. GeoDAR (v1.1) as proportion of ICOLD WRD for each country or territory. (a) By dam count and (b) by reservoir storage capacity. Statistics for Taiwan and Greenland were computed separately from mainland China and Denmark.

Catchment areas of the reservoirs often indicate the stream order of the impounded river, and thus the scales of flow and sediment alterations by the dam. Locating dams with an improved representation of catchment areas, particularly smaller ones, has been increasingly needed by hydrologic modelling and watershed managements (Grill et al., 2019; Lin et al., 2019). To evaluate how GeoDAR spatially resolved WRD in this aspect, we directly used the values of the attribute "catchment area" provided in WRD. As many records in WRD are missing catchment areas, we combined the available values in both WRD and GRanD, and when a dam has catchment areas in both datasets, we preferred the value in GRanD. As reported in Table 5, the subset of WRD georeferenced in GeoDAR v1.1 has a total catchment area of 140 million km^2, which covers 94% of the total catchment area in WRD. The remaining 6% gap was largely closed by the inclusion of the remaining non-WRD dams from GRanD. It is worth mentioning that these statistics do not take into account the dams without documented catchment areas. While it is possible to retrieve catchment boundaries for GeoDAR dams (e.g., using high-resolution DEM as per Mulligan et al. (2020)), acquiring accurate catchment areas of the other WRD dams (which have not been georeferenced) is prohibited due to unknown pour point locations. Therefore, our comparison was only based on the attribute values that are already available. This explains why GeoDAR georeferenced less than half of the WRD records by count but included more than 90% of the total catchment area. Similar to the pattern of reservoir storage capacity, higher proportions of the WRD catchment area covered by GeoDAR are skewed towards the dams with larger catchment areas (Fig. 11a). For example, the number of dams with a catchment area larger than 10 km^2 in GeoDAR equals 88% of that in WRD, and the coverage increases to 95% for the dams with a catchment area larger than 100 km^2.

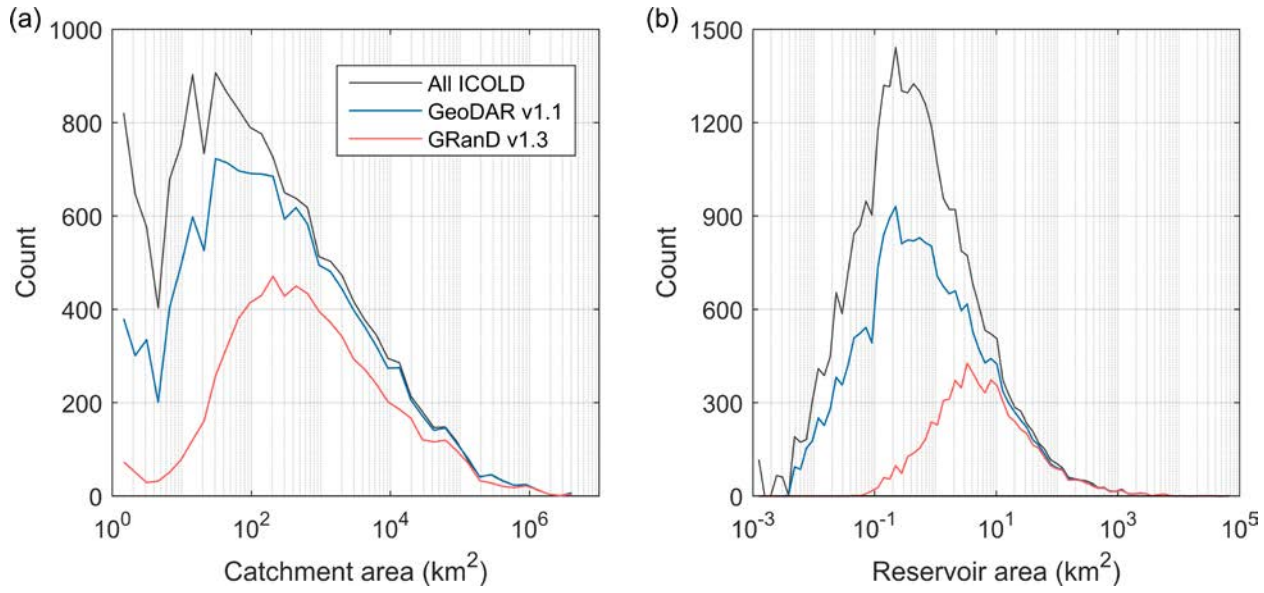


Figure 11. Comparison among GeoDAR, ICOLD WRD, and GRanD by reservoir catchment area and reservoir area. (a) Frequency (count) distributions by reservoir catchment area. Statistics were based on 40 bins between the minimum and maximum catchment areas (i.e., 1 to 4.04 million km^2). (b) Frequency distribution by reservoir area. Statistics are based on 80 bins between the minimum and maximum reservoir areas (i.e., 0.001 to 66,866.7 km^2). All bins are of equal size on a logarithmic scale. Considering that catchment areas are often missing in WRD, a smaller bin size 40 was used to generate smoother distribution curves.

Although GeoDAR does not include reservoir catchment boundaries, it does provide reservoir polygons for 87% of the georeferenced dam points. As reported in Section 3.2, the remaining 13% of the dam points without reservoir polygons, if inferred from their available attribute values, yield a reservoir area that is only 2% of the total reservoir area of all GeoDAR dams. For this reason, we focus on the retrieved reservoir polygons for comparing how GeoDAR v1.1 represents the reservoir areas in the entire ICOLD WRD. Among the 21,515 polygons, 20,718 (96%) are associated with the georeferenced WRD dams. These retrieved WRD reservoirs have a total area of 477 thousand km^2, accounting for 92% of the cumulative reservoir area in WRD (Table 5). After supplementation of the other 797 polygons from GRanD, the total reservoir area reached 496 thousand km^2, equivalent to 96% of the cumulative reservoir area in WRD. Like other attributes, the values of reservoir area are not always available in all WRD records, so our reported coverage percentages are theoretically overestimated. However, if a WRD record is missing its area attribute value but has a retrieved reservoir polygon, we used the area of the reservoir polygon as the *de facto* reservoir area in calculating WRD statistics, and the other WRD records still missing reservoir areas probably contribute a miniscule fraction of the aggregated area. Therefore, we consider our comparison to be overall reasonable. Keeping this limitation in mind, we showed in the distribution curves (Fig. 11b) that the number of GeoDAR reservoir polygons accounts for 68% of all WRD records that have reservoir area values (either

documented or *de facto*), and consistent with the distributions of other attributes, higher coverages for reservoir area tend to occur for larger reservoirs. For example, GeoDAR retrieved 8,263 reservoirs larger than 1 km^2, which account for 80% of those in WRD. The coverage increases to 92% for reservoirs larger than 10 km^2 although the reservoir polygon number decreases to 2,570.

5.2 Improved spatial density over GRanD

While GRanD emphasized dams larger than 100 mcm (or 0.1 km^3), GeoDAR aimed to georeference WRD records which, by definitions, have a minimum storage capacity of 3 mcm or smaller if the dam is higher than 15 m (see Section 1). This reduced storage threshold entailed a substantial increase of the dam quantity in GeoDAR. As compared in Table 5, GeoDAR v1.0, which was generated independently from GRanD, is already more than triple the dam quantity in GRanD (7,320) and accounts for 95% of the total reservoir storage capacity in GRanD (6,881 Gt). With the harmonization with GRanD, the number of dams in GeoDAR v1.1 reaches 339% of that in GRanD, with a total reservoir storage capacity also exceeding 7% of that in GRanD. This comparison suggests that the improvement of GeoDAR is mainly manifested as the increased dam quantity, rather than reservoir storage capacity.

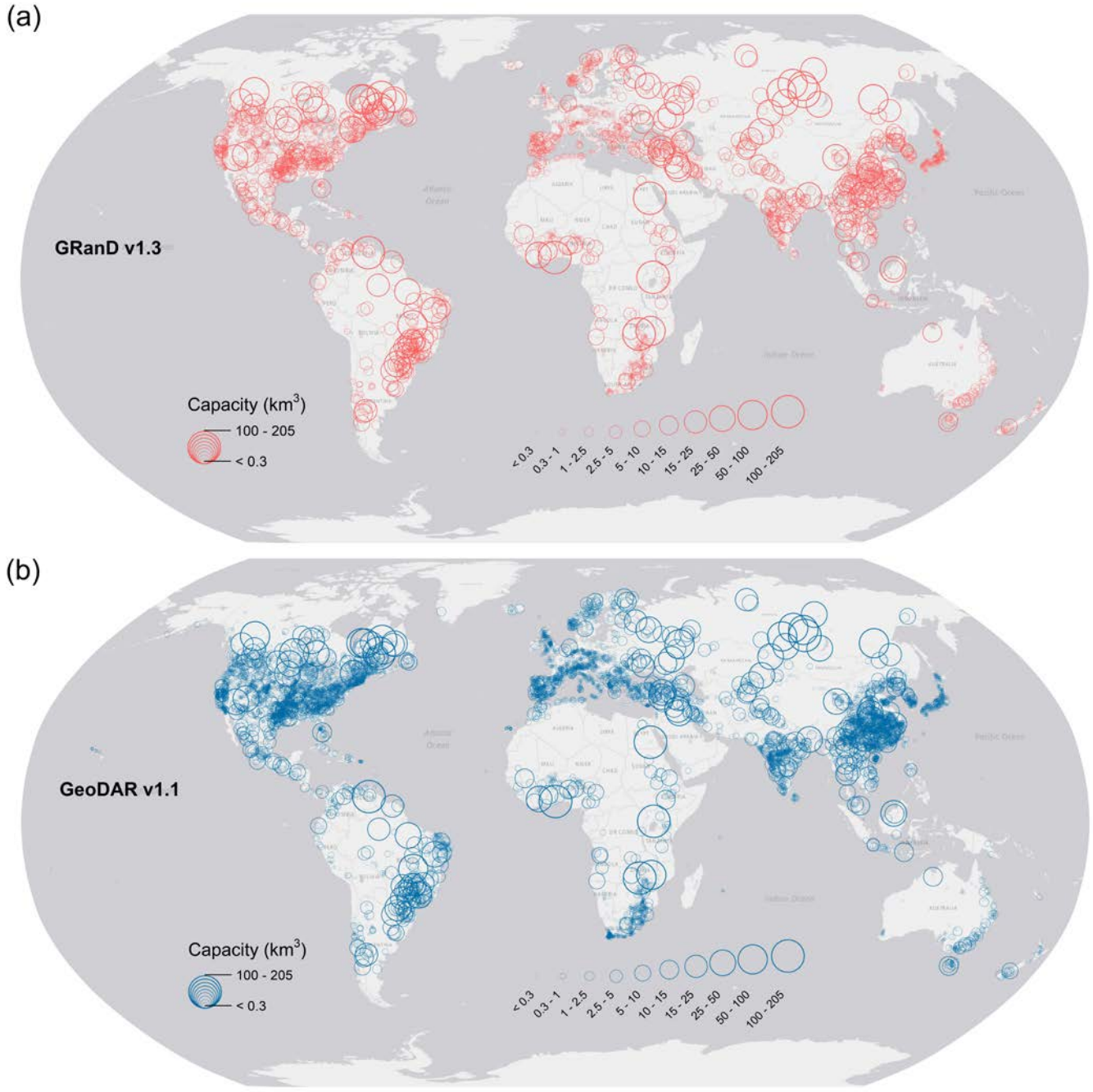
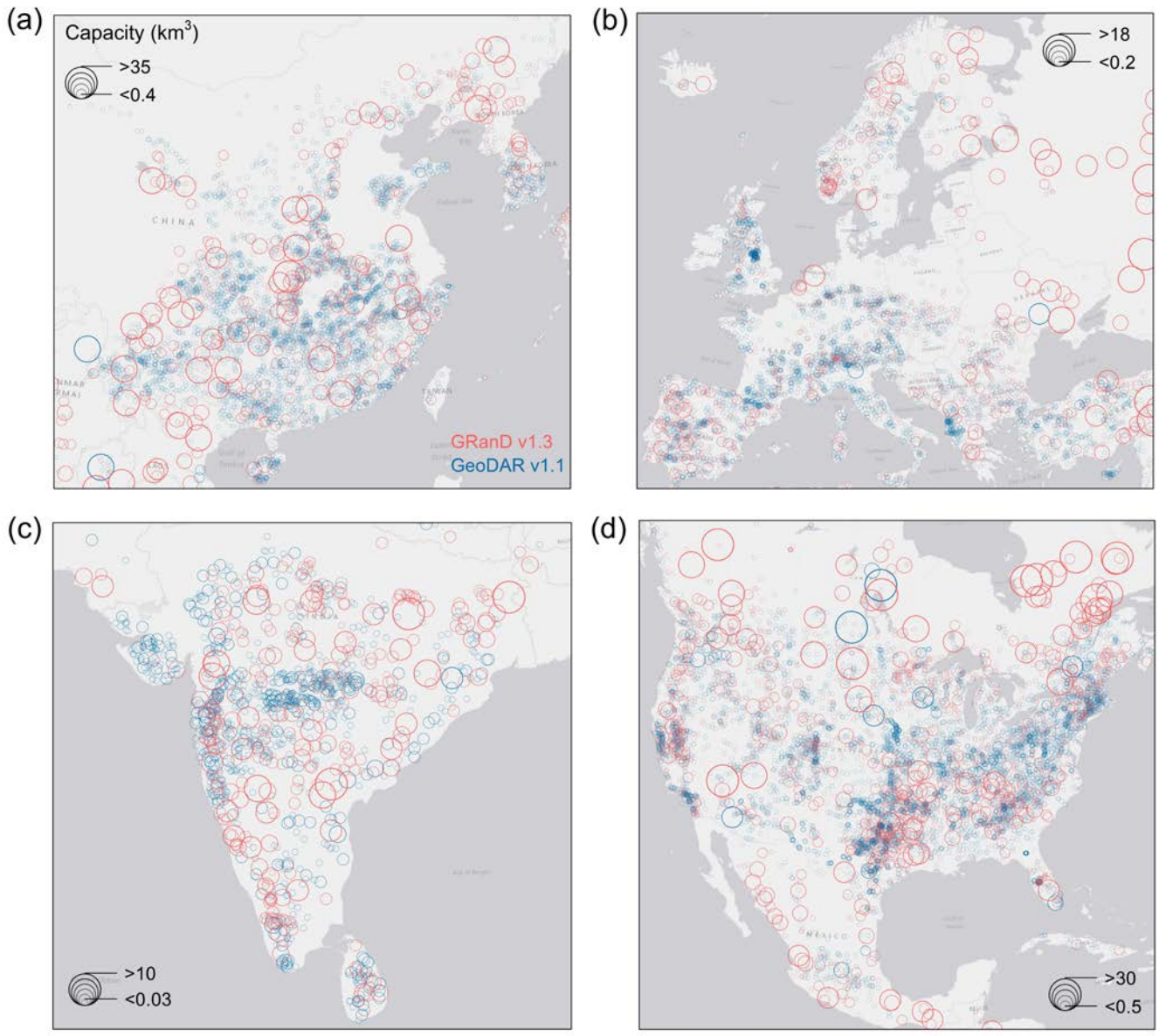


Figure 12. Global distribution of reservoir storage capacities of georeferenced dams. (a) GRanD v1.3 and (b) GeoDAR v1.1. Displayed are 7,312 out of the 7,320 dams in GRanD v1.3 and 24,174 out of the 224,783 dams in GeoDAR v1.1 with documented or estimated reservoir storage capacities.

The increased dam quantity in GeoDAR is manifested as a ubiquitous improvement of the spatial density of smaller dams worldwide (Fig. 12). Since GeoDAR v1.1 has absorbed GRanD v1.3, the global patterns for capacious reservoirs are overall similar between the two datasets. What is noticeably different are the proliferated density of thousands of smaller reservoirs, particularly those beyond the main focus of GRanD (such as smaller than 100 mcm). The substantial increase of smaller dams and reservoirs is corroborated by the distribution curves in Fig. 9a, where the mode storage capacity (i.e., the capacity corresponding to the peak frequency) shifted from about 100 mcm in GRanD to about 3–5 mcm in GeoDAR (both v1.0 and v1.1). The area between the distribution curves is largely explained by the addition of ~16,500 dams smaller than 100 mcm in GeoDAR v1.1 (Fig. 9a), which correspond to a total storage increase of 124 Gt or 95% of the total storage of the dams smaller than 100 mcm in GRanD (Fig. 9b). It is important to note that the added reservoirs in GeoDAR still comply with ICOLD's definition of "large dams" (see Section 1). Although their aggregated storage is limited, these relatively small reservoirs are geographically widespread, meaning that they are locally significant for filling service gaps between more sporadic larger dams. Examples include hundreds of smaller dams/reservoirs that provide irrigation from southern Europe (Fig. 13b) to north-western and central India (Fig. 13c), hydropower and water usage in central and southern China (Fig. 13a), and flood controls across the Mississippi River Basin and southern Texas in the US (Fig. 13d). The sheer number of these added smaller dams and reservoirs accentuate the benefits of an improved knowledge of their spatial locations, such as what GeoDAR offers, for strategizing water and energy managements and assessing fragmentation of the river ecosystems (Belletti et al., 2020; Grill et al., 2019).



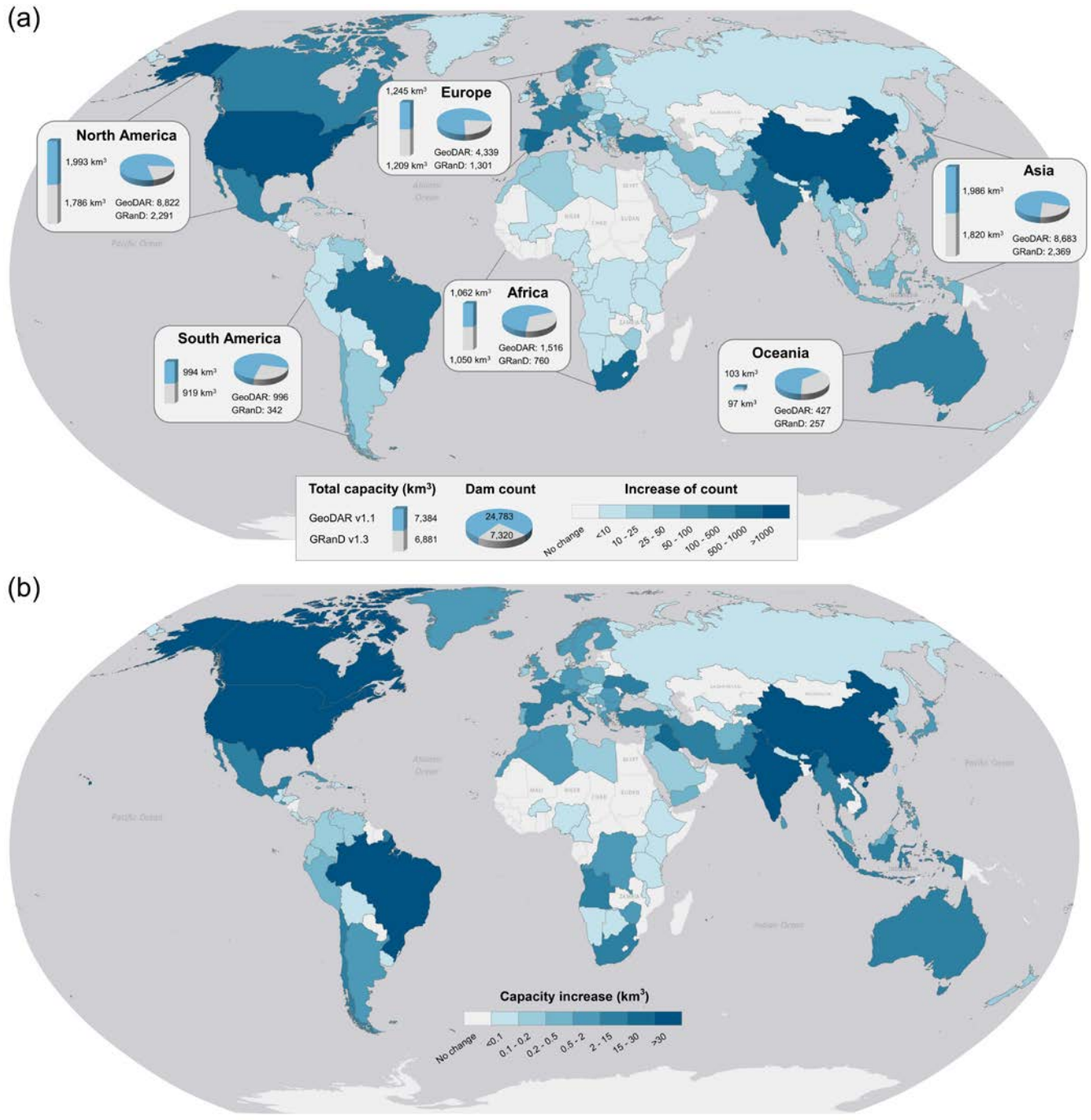
745 **Figure 13.** Regional distributions of reservoir storage capacities in GRanD v1.3 and GeoDAR v1.1. (a) China and its surrounding East and Southeast Asia. (b) Europe. (c) India and its surrounding South Asia. (d) US and its surrounding North America. Graduated symbols for GeoDAR (blue bubbles) are superimposed by symbols for GRanD (red bubbles).

To assist regional applications, we further aggregated the improvements of GeoDAR over GRanD into national scales. As shown in Fig. 14, GeoDAR's improvements in either dam count or reservoir storage capacity pervade more than 120
750 countries occupying 86% of the continental landmass (excluding Antarctica). The increase of dam count occurs in 127 out of

the 155 GeoDAR countries (Fig. 14a). These countries include 18 countries without GRanD records at all (such as Haiti, United Arab Emirates, Yemen, and Bhutan), and the other 109 countries comprise 80% of the 137 countries with GRanD records. There are slightly fewer countries with a confirmed increase of reservoir storage capacity (Fig. 14b) because some of the added WRD records are missing storage capacity values. The number of these countries is 117, including 15 without GRanD records at all.

While GeoDAR's improvements are widespread, the improvement levels are not geographically uniform (Fig. 14). Globally speaking, the spatial patterns of number and capacity increases are overall consistent, with the major hotpots concurring with large or industrialized nations (e.g., US, China, Brazil, India, and European countries) and less impressive increases in smaller, drier, and/or less developed nations (e.g., part of Africa and South America). This is reasonable as bigger and/or more developed nations usually possess a larger quantity of dam infrastructures and thus a greater potential for GeoDAR to improve. 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. The top five countries in terms of dam count increase are the US (an increase of 6,039 or 314%), China (4,352 or474%), India (963 or 290%), South Africa (667 or 248%), and Brazil (575 or 219%) (Supplementary Table S6). These five countries cover nearly three quarters of the global dam count increase (17,463). Similarly, the top five countries in terms of storage capacity increase are the US (123 km^3 or 16%), Canada (73 Gt or 8%), Brazil (66 km^3 or 12%), China (44 km^3 or 7%), and India (33 km^3 or 12%), which together comprise 68% of the global storage capacity increase (503 km^3).

Certain regions with limited increases in dam count, such as the Middle East, Southeast Asia, and southern Africa, show more pronounced improvements in storage capacity. This contrast indicates that, in addition to smaller dams and reservoirs (e.g., <100 mcm), GeoDAR also supplemented GRanD by including more capacious reservoirs. Examples are Dau Tieng Dam in Vietnam (storage capacity 1580 mcm; location 11.323° N, 106.341° E), San Roque Dam in the Philippines (990 mcm; 16.147° N, 120.685° E), Mrica Dam in Indonesia (193 mcm; 7.392° S, 109.605° E), Marib Dam in Yemen (398 mcm; 15.396° N, 45.244° E), and the recently completed Lauca Dam in Angola (5482 mcm; 9.739° S, 15.127° E). GeoDAR also inventoried some large hydroelectric projects that are under construction or consideration. Examples are Bakhtiari Dam in Iran (expected 4,845 mcm; 32.958° N, 48.761° E), Bekhme Dam in Iraq (17,000 mcm; 36.701° N, 44.271° E), Diamer-Bhasha Dam in Pakistan (expected 10,000 mcm; 35.521° N, 73.739° E), and Myitsone Dam in Myanmar (13,282 mcm; 25.691° N, 97.516° E).



780 **Figure 14.** Country-level improvements in GeoDAR v1.1 over GRanD v1.3. (a) Increase of dam count and (b) increase of total reservoir storage capacity for each country or territory. Aggregated statistics for dam count and storage capacity were also compared for each continent. For convenience of comparison, both statistics were displayed on Panel a.

By further aggregating national statistics to each continent (Fig. 14a), the result echoes that GeoDAR's major improvement lies on the quantity or spatial density of the dams, rather than their total reservoir storage capacity. However, this should not overshadow the fact that improvements of both dam count and storage capacity do exist in all continents. As summarized in Fig. 14a, the continental improvement ascends from 173 more dams with a 6 km^3 total capacity in Oceania, to a scale of 6000–7000 more dams with a 100–200 km^3 capacity in North America or Asia. Unfortunately, because the total storage capacity is disproportionally dominated by the largest reservoirs and GRanD has already included most of them, the added storage capacity by GeoDAR relative to what has existed in GRanD appears limited and descends from 9–11% in Asia and North America, 7–8% in Oceania and South America, to 1–3% in Africa and Europe. By contrast, GeoDAR's dam quantity ranges from being almost double that of GRanD in Oceania and Africa, to being triple to quadruple in the other continents.

A derivative benefit of the increased dam quantity is a more complete representation of the regulated watersheds, which is critical to improving discharge estimates. As revealed by the distribution curves in Fig. 11a, GeoDAR improved GRanD in the inclusion of reservoir catchment areas from two aspects. First, the exceedance of the number of reservoir catchments is almost unanimous on all area levels. This corresponds to a total increase of the regulated catchment area by 31,502 km^2 or 27% (Table 5). Second, the increase of reservoir catchments is skewed towards smaller catchments, signifying a more realistic inventory of human water regulations in the basins of lower stream orders or closer to stream headwaters. As shown in the distribution curves (Fig. 11a), the average increasing rate is augmented from about 30% for catchments larger than 1000 km^2, 80% for catchments between 10 and 1000 km^2, to more than 600% for those smaller than 10 km^2. The mode of catchment areas decreases from about 200–400 km^2 in GRanD to 30–100 km^2 in GeoDAR, with the latter much closer to the mode of the entire WRD (15–50 km^2). As a result, the number of dams with a catchment size smaller than 25 km^2, for example, which is the channelization threshold for the high-resolution MERIT Basins hydrography dataset (Lin et al., 2019; Yamazaki et al., 2017)), is 3,570 or 27% in GeoDAR in comparison to 695 or 10% in GRanD. These small-catchment dams, once integrated into river networks, may substantially improve the performance of routing models. Consistent with our comparison with WRD (Section 5.1), these statistics are only based on the records with valid catchment areas. Considering that missing values more likely occur to dams with smaller catchments, our reported improvement could be theoretically conservative.

The increased dam count in GeoDAR also enabled the retrieval of surface extents of another 14,000 or so smaller reservoirs (Fig. 7). These added reservoir polygons have an average size of 1.4 km^2 in comparison to 65 km^3 in GRanD. They aggregate to a total area of 19,880 km^2, a scale comparable to 30 Lake Meads. Although this area increase may appear substantial, it only expanded the global reservoir area in GRanD by a marginal proportion of 4%. Similar to the pattern of storage capacities, reservoir areas follow a quasi-Pareto distribution, meaning that smaller reservoirs tend to dominate the population (or number) whereas larger reservoirs dominate the area and storage. This explains why the increase of relative area is small, but the increase of absolute quantity is double that of the entire reservoir polygons in GRanD. For example,

95% of the total reservoir area in GeoDAR comes from only 12% of the reservoir polygons larger than 10 km^2, and about 90% of these large reservoirs are already included by GRanD (Fig. 11b). This pattern again suggests that the core value of GeoDAR is not to augment the global scale of reservoir area or storage, but to amplify the local details of smaller dams and reservoirs. Owing to the added details, the mode of reservoir area is on the order of 1–10 km^2 in GRanD but was refined by one order of magnitude to 0.1–1 km^2 in GeoDAR.

If we group the global dams by their documented main purpose, we observe in Fig. 15 that GeoDAR improved GRanD unanimously in both dam count and storage capacity for all main purposes (Fig. 15). For the same reason as explained above (i.e., the added reservoirs are small), the increases of dam count appear more prominent than those of storage capacity, and the increases of storage capacity from GRanD to GeoDAR are overall more evident than those from GeoDAR to ICOLD WRD. The exception is the dams with “others” or “unknown” purposes whose total storage capacity in GeoDAR is lower. This is because when GRanD and WRD records conflict with each other in the GeoDAR harmonization process, the attribute values in GRanD took precedence only if they are available or valid (“others” or “unknown” was considered as invalid reservoir purpose). Assuming that reservoir operations vary by purpose, this unanimous improvement of the spatial inventory for all reservoir purposes, in conjunction with satellite-observed water budget variations, can help us better generalize reservoir operation rules which are critical to improving water managements.

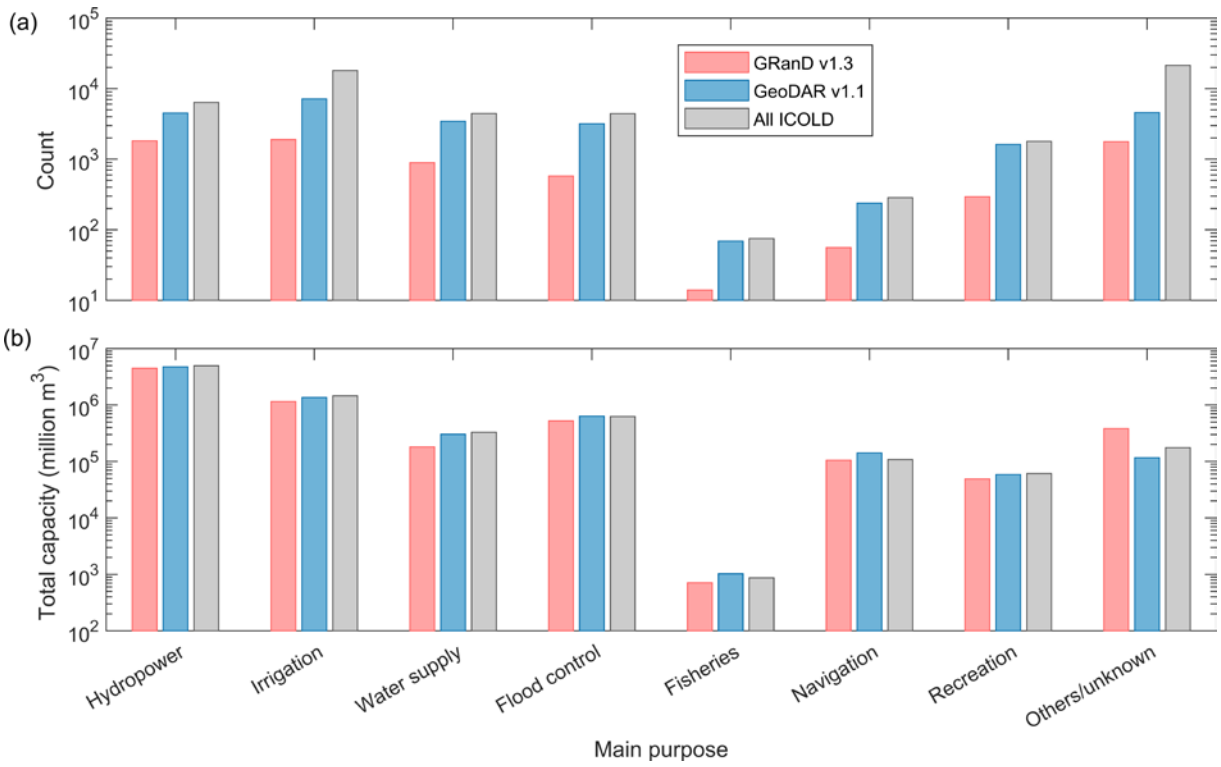


Figure 15. Comparison among GRanD v1.3, GeoDAR v1.1, and ICOLD WRD by dam/reservoir purpose. (a) Dam counts and (b) total reservoir storage capacities for each main purpose. Dam purposes are based on attribute values provided in WRD and GRanD. For a dam with multiple purposes, its "main purpose" was considered as the one with the highest order of priority. The main purpose in GRanD took precedence if it differs from that in WRD.

5.3 Spatially complementary to GOODD

The recently published GOODD (V1) dataset (Mulligan et al., 2020) includes 38,667 dam points in the world, which were digitized by scanning through Google Earth imagery with supports of regional inventories and the Shuttle Radar Topography Mission Water Body Dataset (SWBD, 2005). Despite lacking essential attributes, GOODD is thus far the most comprehensive global inventory of dam locations and catchments. The digitization was performed during 2007 to 2011 and was later updated in 2016. This means that reservoirs postdating 2016 were not yet included in the dataset. The completeness and accuracy of GOODD also depend on the sizes of the dams or reservoirs. According to Mulligan et al. (2020), the resolution and quality of available Google Earth imagery during the digitization period were low in some parts of the world (such as China), and an experiment in the US showed that detectable dams and reservoirs from low resolution imagery (e.g., Landsat Geocover 2000) may require the reservoir length greater than 500 m and the dam width greater than 150 m. These minimum size criteria do not necessarily overlap with those of ICOLD WRD which instead emphasize the reservoir storage capacity and dam height (see Section 1).

Because of these digitizing limitations and criterion difference, the dam points in GeoDAR are spatially complementary to, rather than always duplicated by, those in GOODD across many regions. Figure 16 identified four examples in Cerrado Brazil, northern China, southwestern France, and northern Pakistan, where a large proportion of the GeoDAR dams were not digitized by GOODD. Some of the dams that only appear in GeoDAR also comply with the minimum size criteria of GOODD, and examples are those enlarged in the right panels except the Duber Khwar Dam in Pakistan (35.119° N, 72.927° E; Fig. 16j) which was completed more recently in 2014. Since the area of the Duber Khwar Reservoir (about 0.05 km^2) is smaller than the resolution of HydroLAKES (0.1 km^2) and the dam completion year overlaps with the image acquisition period of the UCLA Circa 2015 Lake Inventory (from May 2013 to August 2015 (Sheng et al., 2016)), GeoDAR georeferenced the dam point but did not successfully retrieve the reservoir polygon.

To approximate how GeoDAR and GOODD complement each other globally, we intersected both dam datasets with the 30-m-resolution UCLA Circa 2015 Lake Inventory. As a result of manual snapping to the 30-arc second HydroSHEDS streamflow network (Lehner et al., 2008), some of the points in GOODD ended up having substantial geographic offsets from the actual locations. For a pilot experiment, we applied a 1-km tolerance (about 30-arc-second on the equator) when intersecting the UCLA lake inventory with GOODD, and kept a 500-m tolerance as used in Section 2.5 for intersecting the lake inventory with GeoDAR. The result shows that among the 55,000 or so water bodies that intersect either datasets, 80% intersect with GOODD and the other 20% with GeoDAR alone. These statistics imply that GeoDAR may have an ability to expand the number of dams in GOODD by roughly 25% (i.e., 20% divided by 80%). Since we applied a larger tolerance for GOODD, this estimated expansion by GeoDAR is likely conservative (considering that the number of GOODD-intersecting reservoirs may be overestimated). If a 500-m tolerance is used for both intersections, the expansion by GeoDAR will increase to roughly 45%. In addition to the expanded spatial coverage, GeoDAR indexed each georeferenced dam point to a WRD and/or GRanD record and thus enabled access to multiple attributes, whereas GOODD carries no attribute information except the delineated reservoir catchments. These regional and global comparisons suggest that, even just with the geometric dam points, GeoDAR is not a simple replication of GOODD, but instead complements GOODD for an improved spatial coverage and density of global dams.

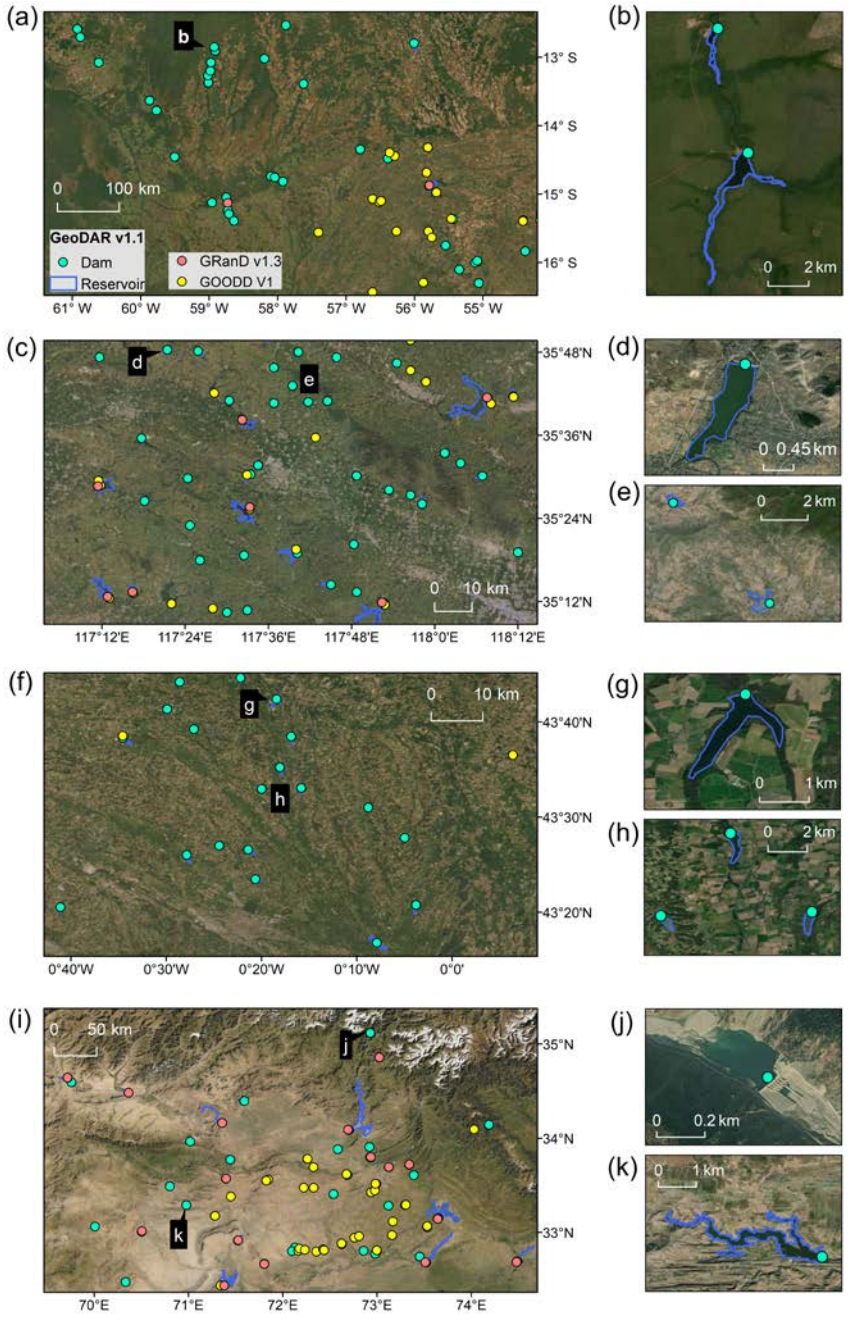


Figure 16. Comparisons between GRanD v1.3, GOODD V1, and GeoDAR v1.1 in selected regions of the world. (a)-(b) Cerrado, Brazil (Mato Grasso State). (c)-(e) Northern China (Shandong Province). (f)-(h) Southwestern France (Aquitaine and Midi-Pyrenees). (i)-(k) Northern Pakistan (northern highlands and foothills). GRanD points (red) are placed on top of GOODD (green) which is placed on top of GeoDAR (yellow). Background image source: Esri imagery base map.

6 Data availability

GeoDAR v1.0 (dam points) and v1.1 (both dam points and reservoir polygons) are available for download from the Zenodo repository https://doi.org/10.5281/zenodo.6163413. The dam points are stored in both csv and shapefile formats, and the reservoir polygons are provided in shapefile. Their attributes and values are described in Table 3 as well as in the repository website. The data usage information is described in Section 3.3. Other citation courtesy and disclaimer information are given in the Disclaimer section and the repository website. All released datasets and information are available under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license (https://creativecommons.org/licenses/by/4.0). Users who would like to link GeoDAR records to the proprietary WRD attributes they have purchased in advance from ICOLD should contact the corresponding author.

7 Summary and applications

We have produced a comprehensive and spatially resolved dam and reservoir dataset, GeoDAR, which complementarily improved the existing global inventories of large dams. We demonstrated that the production of GeoDAR is not a direct compilation or collation of existing dam datasets. Instead, it involved a first-known effort to georeference ICOLD WRD. This was jointly enabled by geo-matching (or table-associating) multi-source regional registers and geocoding descriptive attributes through the Google Maps API. This georeferencing effort resulted in GeoDAR v1.0 which contains 22,560 spatially resolved dam points, each associated with a WRD record, with an overall accuracy of 95%. Each of the georeferenced records was also labelled with a QA score, providing users a reference to the qualities of individual dam locations. Our georeferencing process and accuracy validation, as we have elaborated in substantive detail, have important methodological values for future expansions of spatial dam inventories using similar approaches, such as Geo-Wiki and OpenStreetMap.

To further ensure the optimal inclusion of the world's largest dams, we harmonized the georeferenced WRD (or GeoDAR v1.0) carefully with GRanD v1.3. Using the harmonized dam points as spatial identifiers, most of their reservoir boundaries were then retrieved from high-resolution water body datasets. This ICOLD-GRanD harmonization and the subsequent reservoir retrieval resulted in GeoDAR v1.1, our end product, which holds 24,783 dam points (including 23,974 linked to WRD) and 21,515 reservoir polygons. This product spatially resolved 44% of the entire ICOLD WRD by dam count and more than 90% by reservoir storage capacity. Since most of the world's largest reservoirs (e.g., >0.1 km^3) are already included in GRanD, GeoDAR adds limited improvements (by 4–27%) to the total reservoir area, storage capacity, and catchment area. However, by including many smaller dams particularly in lower and middle latitudes, GeoDAR is triple the size of GRanD in terms of dam and reservoir quantity. For this reason, one of the major improvements of GeoDAR is its unparalleled ability to capture relatively small dams, or in other words, to enhance the spatial detail of global dam and reservoir distributions.

Besides an improved quantity and spatial detail, another unique value of GeoDAR is its capability of bridging the locations of dams to a broad suite of attributes that are essential to scientific applications. A standing dilemma of existing global dam datasets is the divergence between the focus on dam quantity or spatial detail and the provision of detailed attributes for a limited dam quantity. This dilemma was partially ameliorated by GeoDAR because its georeferenced dams and reservoirs were explicitly indexed to WRD and/or GRanD records where many attributes are available. Since the original WRD is not georeferenced, our perception was that the task of georeferencing WRD to enable a spatially explicit application of the attribute information, even at regional scales, may fell on individual users. To avoid the duplication of efforts and to facilitate scientific applications, we performed this comprehensive georeferencing on the entirety of ICOLD WRD as thoroughly as possible, and hereby released the resultant dam coordinates and reservoir polygons to the public as part of GeoDAR. We would like to reiterate the disclaimer that GeoDAR does not directly contain, and neither do we intend to release, the original WRD attribute data which are proprietary to ICOLD. In other words, the association between GeoDAR IDs and WRD IDs exist but were purposefully encrypted. However, if individual users need GeoDAR records to be linked to the WRD attributes that they already purchased from ICOLD, we can be contacted and on a case-by-case basis, we may provide this assistance given that the users agree not to release the decryption key or the proprietary WRD attributes.

We envision that GeoDAR, with its enhanced spatial density and extended accessibility to essential attributes, will benefit a wide spectrum of disciplines and applications. It is worth noting that although most dams in GeoDAR are smaller than those in GRanD or AQUASTAT, they are still compliant with ICOLD's size criteria which exclude countless tiny on-farm reservoirs and water storage tanks. Nevertheless, we have suggested from regional examples that GeoDAR partially complements some of the most extensive global dam inventories such as GOODD, despite GOODD owning a larger number of dams. In this sense, even just with the 25,000 or so geometric dam points, GeoDAR contributes yet another fundamental extension to global water infrastructure databases. If these dam points are rectified to high-resolution hydrographic networks (such as MERIT Hydro (Lin et al., 2021; Yamazaki et al., 2019)), GeoDAR, together with other existing dam and barrier datasets, can help refine our understanding of how human water infrastructure fragmented global rivers and their ecosystems (Belletti et al., 2020; Grill et al., 2019; Yang et al., 2022), especially with a more exhaustive inclusion of smaller and/or headwater catchments.

Alongside the detailed dam points, GeoDAR's reservoir boundaries provide thus far the most comprehensive global base maps for assessing reservoir dynamics and the impacts of human water regulation. In combination with the expanding constellation of satellite sensors (e.g., ICESat-2, Sentinel-6, and the forthcoming SWOT), this high-resolution base map will, for instance, enable a more complete and accurate monitoring of water storage variation and surface evaporation in global reservoirs (Biancamaria et al., 2016; Chen et al., 2021; Cretaux et al., 2016; Zhao and Gao, 2019a). Tracking the spatiotemporal balance between reservoir water storage and evaporative loss will help strategize regional water managements under a warming climate (Cretaux et al., 2015). Since our knowledge and understanding improves as

observations increase, the observed water storage dynamics for an increased quantity of reservoirs will inevitably entail a more realistic generalization of the reservoir operation rules. This is particularly true if the attribute information such as reservoir purpose and storage capacity are also utilized. Considering that small but widespread reservoirs have a strong cumulative impact on discharge (Habets et al., 2018; Lin et al., 2019), the improved operation rules and the fine details of reservoir storage changes will benefit discharge estimations from hydrological models. From another perspective, GeoDAR's reservoir polygons can also help refine surface water typology, either by directly using them to mask artificial impoundments from natural lakes, or by expanding the training pool to enhance machine learning algorithms so that additional reservoirs can be detected (Fang et al., 2019). A refined water typology map will, in turn, assist other analysis tools in improving our assessments of how human footprints alter surface hydrology and its related biodiversity and ecosystem health.

8 Code availability

Python scripts for geo-matching, geocoding, and reservoir assignment are publicly available at https://github.com/surf-hydro/georeferencing-ICOLD-dams-and-reservoirs. We request users who adapt or use the scripts to cite Wang et al. (2021).

9 Author contribution

JW: Conceptualization, Data curation, Data harmonization, Formal analysis, Funding acquisition, Investigation, Methodology, Programming, Project administration, Quality assurance, Quality control, Supervision, Validation, Visualization, Writing – original draft preparation, Writing – revision. BAW: Data curation, Formal Analysis, Investigation, Methodology, Programming, Visualization, Writing – original draft preparation, Writing – review and editing. FY: Data curation, Methodology, Quality control, Writing – review and editing. CQ: Methodology, Quality control, Supervision, Validation, Writing – review and editing. MD: Quality control, Validation, Writing – review. ASM: Quality control, Validation, Writing – review. JZ: Quality control, Validation. CF: Quality control, Validation. JMM: Validation, Writing – review and editing. SS: Methodology, Writing – review and editing. YS: Data curation, Methodology, Supervision, Writing – review and editing. GHA: Methodology, Supervision, Writing – review and editing. JFC: Methodology, Supervision, Writing – review and editing. YW: Methodology, Supervision, Writing – review and editing.

10 Competing interests

The authors declare no conflict of interest.

11 Disclaimer

GeoDAR v1.0 and v1.1 contain knowledge derived from ICOLD WRD (https://www.icold-cigb.org/GB/world_register/acknowledgements_wrd.asp) but release no original values of the proprietary WRD attributes (except the storage capacities of a few large dams used to verify and correct Wada et al. (2017); see Table S4). The production and dissemination of GeoDAR (spatial features) abide by ICOLD's legal policies (https://www.icold-cigb.org/GB/legal.asp) and were approved by the Central Office of ICOLD. GeoDAR v1.0 represents an initial effort of georeferencing WRD at a global scale, and the resultant dam distribution may be geographically skewed and thus may not reflect the distribution of all WRD records. The authors are not responsible for any consequence arising from this limitation. GeoDAR v1.1 absorbed the spatial features (i.e., dam point coordinates and most of the reservoir polygons) in GRanD v1.3. To acknowledge the originality of GRanD, we request users to cite Lehner et al. (2011) if they only use the subset of GeoDAR v1.1 from GRanD alone. If GRanD is used together with our corrected spatial coordinates (Supplementary Table S5), we recommend users citing this paper as well. The source of each spatial feature in GeoDAR v1.1 is specified in the attributes "har_src" and "pnt_src" for dam points and the attribute "plg_src" for reservoir polygons (see Table 3). For any questions about data citation, please contact the corresponding author JW. Authors of this paper claim no responsibility or liability for any consequences related to the use, citation, or dissemination of GeoDAR.

12 Acknowledgements

The work was in part supported by NASA Surface Water and Ocean Topography (SWOT) Grant (#80NSSC20K1143) to JW and Kansas State University faculty start-up fund to JW. We would like to acknowledge ICOLD for providing WRD and the Central Office of ICOLD for informing data dissemination policies and for allowing us to release the position information of WRD we georeferenced. We thank Aote Xin at Kansas State University for assisting in data harmonization, quality control and validation, and for providing helpful comments on the manuscript. We thank Dr. Yao Li at Texas A&M University for informing incomplete reservoir polygons, and Elizabeth M. Prior at Virginia Tech for informing some of the duplicate dam points in the US. The authors are also grateful to Dr. Bernhard Lehner at McGill University for his constructive suggestions and comments on data curation, usage, and dissemination. We also acknowledge Google Maps Platform (https://cloud.google.com/maps-platform) for providing the geocoding API.

13 References

Allen, G. H. and Pavelsky, T. M.: Global extent of rivers and streams, Science, 361, 585-587, https://doi.org/10.1126/science.aat0636, 2018.

Belletti, B., Leaniz, C. G. d., Jones, J., Bizzi, S., Börger, L., Segura, G., Castelletti, A., Bund, W. v. d., Aarestrup, K., Barry, J., Belka, K., Berkhuysen, A., Birnie-Gauvin, K., Bussettini, M., Carolli, M., Consuegra, S., Dopico, E., Feierfeil, T.,

Fernández, S., Garrido, P. F., Garcia-Vazquez, E., Garrido, S., Giannico, G., Gough, P., Jepsen, N., Jones, P. E., Kemp, P., Kerr, J., King, J., Łapińska, M., Lázaro, G., Lucas, M. C., Marcello, L., Martin, P., McGinnity, P., O'Hanley, J., Amo, R. O. d., Parasiewicz, P., Pusch, M., Rincon, G., Rodriguez, C., Royte, J., Schneider, C. T., Tummers, J. S., Vallesi, S., Vowles, A., Verspoor, E., Wanningen, H., Wantzen, K. M., Wildman, L., and Zalewski, M.: More than one million barriers fragment Europe's rivers, Nature, 588, 436-441, https://doi.org/10.1038/s41586-020-3005-2, 2020.

Biancamaria, S., Lettenmaier, D. P., and Pavelsky, T. M.: The SWOT mission and its capabilities for land hydrology, Surv. Geophys., 37, 307-337, https://doi.org/10.1007/s10712-015-9346-y, 2016.

Biemans, H., Haddeland, I., Kabat, P., Ludwig, F., Hutjes, R. W. A., Heinke, J., von Bloh, W., and Gerten, D.: Impact of reservoirs on river discharge and irrigation water supply during the 20th century, Water Resour. Res., 47, W03509, https://doi.org/10.1029/2009WR008929, 2011.

Boulange, J., Hanasaki, N., Yamazaki, D., and Pokhrel, Y.: Role of dams in reducing global flood exposure under climate change, Nat. Commun., 12, 417, https://doi.org/10.1038/s41467-020-20704-0, 2021.

Busker, T., de Roo, A., Gelati, E., Schwatke, C., Adamovic, M., Bisselink, B., Pekel, J. F., and Cottam, A.: A global lake and reservoir volume analysis using a surface water dataset and satellite altimetry, Hydrol. Earth Syst. Sci., 23, 669-690, https://doi.org/10.5194/hess-23-669-2019, 2019.

Carpenter, S. R., Stanley, E. H., and Vander Zanden, M. J.: State of the world's freshwater ecosystems: physical, chemical, and biological changes, Annu. Rev. Environ. Resour., 36, 75-99, https://doi.org/10.1146/annurev-environ-021810-094524, 2011.

Chao, B. F., Wu, Y. H., and Li, Y. S.: Impact of artificial reservoir water impoundment on global sea level, Science, 320, 212-214, https://doi.org/10.1126/science.1154580, 2008.

Chen, T., Song, C., Ke, L., Wang, J., Liu, K., and Wu, Q.: Estimating seasonal water budgets in global lakes by using multi-source remote sensing measurements, Journal of Hydrology, 593, 125781, https://doi.org/10.1016/j.jhydrol.2020.125781, 2021.

Crétaux, J. F., Abarca-del-Rio, R., Berge-Nguyen, M., Arsen, A., Drolon, V., Clos, G., and Maisongrande, P.: Lake volume monitoring from space, Surv. Geophys., 37, 269-305, https://doi.org/10.1007/s10712-016-9362-6, 2016.

Crétaux, J. F., Biancamaria, S., Arsen, A., Berge-Nguyen, M., and Becker, M.: Global surveys of reservoirs and lakes from satellites and regional application to the Syrdarya river basin, Environ. Res. Lett., 10, 015002, http://dx.doi.org/10.1088/1748-9326/10/1/015002, 2015.

Crétaux, J. F., Jelinski, W., Calmant, S., Kouraev, A., Vuglinski, V., Berge-Nguyen, M., Gennero, M. C., Nino, F., Del Rio, R. A., Cazenave, A., and Maisongrande, P.: SOLS: A lake database to monitor in the near real time water level and storage variations from remote sensing data, Adv. Space. Res., 47, 1497-1507, https://doi.org/10.1016/j.asr.2011.01.004, 2011.

Dams in Japan, Japan Dam Foundation (JDF): http://damnet.or.jp/Dambinran/binran/TopIndex_en.html, last access: May 2021.

Degu, A. M., Hossain, F., Niyogi, D., Pielke, R., Shepherd, J. M., Voisin, N., and Chronis, T.: The influence of large dams on surrounding climate and precipitation patterns, Geophys. Res. Lett., 38, L04405, https://doi.org/10.1029/2010GL046482, 2011.

Department of Water and Sanitation (DWS) of South Africa: List of Registered Dams (LRD) [data set], http://www.dwaf.gov.za/DSO/Publications.aspx, 2019.

Döll, P., Fiedler, K., and Zhang, J.: Global-scale analysis of river flow alterations due to water withdrawals and reservoirs, Hydrol. Earth Syst. Sci., 13, 2413-2432, https://doi.org/10.5194/hess-13-2413-2009, 2009.

Fang, W., Wang, C., Chen, X., Wan, W., Li, H., Zhu, S., Fang, Y., Liu, B., and Hong, Y.: Recognizing global reservoirs from Landsat 8 images: a deep learning approach, IEEE J. Sel. Top. Appl. Earth. Obs. Remote Sens., 12, 3701-3701, https://doi.org/10.1109/JSTARS.2019.2929601, 2019.

Gao, H., Birkett, C., and Lettenmaier, D. P.: Global monitoring of large reservoir storage from satellite remote sensing, Water Resour. Res., 48, W09504, https://doi.org/10.1029/2012WR012063, 2012.

Grill, G., Lehner, B., Thieme, M., Geenen, B., Tickner, D., Antonelli, F., Babu, S., Borrelli, P., Cheng, L., Crochetiere, H., Macedo, H. E., Filgueiras, R., Goichot, M., Higgins, J., Hogan, Z., Lip, B., McClain, M. E., Meng, J., Mulligan, M., Nilsson, C., Olden, J. D., Opperman, J. J., Petry, P., Liermann, C. R., Saenz, L., Salinas-Rodriguez, S., Schelle, P., Schmitt, R. J. P., Snider, J., Tan, F., Tockner, K., Valdujo, P. H., van Soesbergen, A., and Zarfl, C.: Mapping the world's free-flowing rivers, Nature, 569, 215-221, https://doi.org/10.1038/s41586-019-1111-9, 2019.

Goteti, G. and Stachelek J.: Dams in the United States from the National Inventory of Dams, R package version 0.2 [data set], https://www.rdocumentation.org/packages/dams/versions/0.2, 2016.

Habets, F., Molenat, J., Carluer, N., Douez, O., and Leenhardt, D.: The cumulative impacts of small reservoirs on hydrology: a review, Sci. Total Environ., 643, 850-867, https://doi.org/10.1016/j.scitotenv.2018.06.188, 2018.

Latrubesse, E. M., Arima, E. Y., Dunne, T., Park, E., Baker, V. R., d'Horta, F. M., Wight, C., Wittmann, F., Zuanon, J., Baker, P. A., Ribas, C. C., Norgaard, R. B., Filizola, N., Ansar, A., Flyvbjerg, B., and Stevaux, J. C.: Damming the rivers of the Amazon basin, Nature, 546, 363-369, https://doi.org/10.1038/nature22333, 2017.

Lehner, B., Liermann, C. R., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., Frenken, K., Magome, J., Nilsson, C., Robertson, J. C., Rodel, R., Sindorf, N., and Wisser, D.: High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management, Front. Ecol. Environ., 9, 494-502, https://doi.org/10.1890/100125, 2011.

Lehner, B., Verdin, K., and Jarvis, A.: New global hydrography derived from spaceborne elevation data, Eos, Transactions, American Geophysical Union, 89, 93-104, http://doi.org/10.1029/2008eo100001, 2008.

Li, B., Yan, Q., and Zhang, L.: Flood monitoring and analysis over the middle reaches of Yangtze River basin using MODIS time-series imagery, in: 2011 IEEE International Geoscience and Remote Sensing Symposium, Vancouver, British Columbia, Canada, 24-29 July 2011, 807-810, https://doi.org/10.1109/IGARSS.2011.6049253, 2011.

Li, Y., Gao, H., Zhao, G., and Tseng, K. H.: A high-resolution bathymetry dataset for global reservoirs using multi-source satellite imagery and altimetry, Remote Sens. Environ., 244, 111831, https://doi.org/10.1016/j.rse.2020.111831, 2020.

Lin, P., Pan, M., Wood, E. F., Yamazaki, D., and Allen, G. H.: A new vector-based global river network dataset accounting for variable drainage density, Sci. Data, 8, 28, https://doi.org/10.1038/s41597-021-00819-9, 2021.

Lin, P., Pan, M., Beck, H. E., Yang, Y., Yamazaki, D., Frasson, R., David, C. H., Durand, M., Pavelsky, T. M., Allen, G. H., Gleason, C. J., and Wood, E. F.: Global reconstruction of naturalized river flows at 2.94 million reaches, Water Resour. Res., 55, 6499-6516, https://doi.org/10.1029/2019WR025287, 2019.

Liu, K., Song, C., Wang, J., Ke, L., Zhu, Y., Zhu, J., Ma, R., and Luo, Z.: Remote sensing-based modeling of the bathymetry and water storage for channel-type reservoirs worldwide, Water Resour. Res., 56, e2020WR027147, https://doi.org/10.1029/2020WR027147, 2020.

Lyons, E. A. and Sheng, Y.: LakeTime: Automated seasonal scene selection for global lake mapping using Landsat ETM+ and OLI, Remote Sensing, 10, 54, https://doi.org/10.3390/rs10010054, 2018.

Mady, B., Lehmann, P., Gorelick, S. M., and Or, D.: Distribution of small seasonal reservoirs in semi-arid regions and associated evaporative losses, Environ. Res Commun., 2, 061002, https://doi.org/10.1088/2515-7620/ab92af, 2020.

Managing Aquatic ecosystems and water Resources under multiple Stress project (MARS): MARS GeoDatabase (MARSgeoDB) version 2 [data set], http://www.mars-project.eu/index.php/databases.html, 2017.

Map World (Tianditu), National Platform for Common Geospatial Information Services (NPCGIS): https://map.tianditu.gov.cn, last access: July 2021.

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., and Schmitt, O.: Estimating the volume and age of water stored in global lakes using a geo-statistical approach, Nat. Commun., 7, 13603, https://doi.org/10.1038/ncomms13603, 2016.

Mulligan, M., van Soesbergen, A., and Saenz, L.: GOODD, a global dataset of more than 38,000 georeferenced dams, Sci. Data, 7, 31, https://doi.org/10.1038/s41597-020-0362-5, 2020.

National Register of Large Dams (NRLD), Central Water Commission (CWC), Government of India, New Delhi, 281 pp., 2019.

Natural Resources Canada (NRC): CanVec 1M Man-Made Features - Dam version 1.0 [data set], http://geogratis.gc.ca/api/en/nrcan-rncan/ess-sst/0c78d7fe-100b-5937-b74e-7590a03a6244.html, 2017.

Nilsson, C. and Berggren, K.: Alterations of riparian ecosystems caused by river regulation, Bioscience, 50, 783-792, https://doi.org/10.1641/0006-3568(2000)050[0783:AORECB]2.0.CO;2, 2000.

Open Development Cambodia (ODC): Hydropower dams 1993-2014 [data set], https://data.opendevelopmentmekong.net/en/dataset/hydropower-2009-2014, 2015.

Open Development Myanmar (ODM): Myanmar Dams [data set], https://data.opendevelopmentmekong.net/en/dataset/myanmar-dams, 2018.

Paredes-Beltran, B., Sordo-Ward, A., and Garrote, L.: Dataset of Georeferenced Dams in South America (DDSA), Earth Syst. Sci. Data, 13, 213-229, https://doi.org/10.5194/essd-13-213-2021, 2021.

Pekel, J. F., Cottam, A., Gorelick, N., and Belward, A. S.: High-resolution mapping of global surface water and its long-term changes, Nature, 540, 418-422, https://doi.org/10.1038/nature20584, 2016.

Schwatke, C., Dettmering, D., Bosch, W., and Seitz, F.: DAHITI - an innovative approach for estimating water level time series over inland waters using multi-mission satellite altimetry, Hydrol. Earth Syst. Sci., 19, 4345-4364, https://doi.org/10.5194/hess-19-4345-2015, 2015.

Sheng, Y., Song, C., Wang, J., Lyons, E. A., Knox, B. R., Cox, J. S., and Gao, F.: Representative lake water extent mapping at continental scales using multi-temporal Landsat-8 imagery, Remote Sens. Environ., 185, 129-141, https://doi.org/10.1016/j.rse.2015.12.041, 2016.

Shin, S., Pokhrel, Y., and Miguez-Macho, G.: High-resolution modeling of reservoir release and storage dynamics at the continental scale, Water Resour. Res., 55, 787-810, https://doi.org/10.1029/2018WR023025, 2019.

Shuttle Radar Topography Mission Water Body Data set (SWBD): http://www2.jpl.nasa.gov/srtm, last access 2014.

Sistema Nacional de Informações sobre Segurança de Barragens (SNISB, Brazilian National Dam Safety Information System): Relatório de Segurança de Barragens 2017 (Dams Safety Report 2017) [data set], http://www.snisb.gov.br/portal/snisb/relatorio-anual-de-seguranca-de-barragem/2017, 2017.

Tilt, B., Braun, Y., and He, D.: Social impacts of large dam projects: A comparison of international case studies and implications for best practice, Journal of Environmental Management, 90, S249-S257, 2009.

Tobler, W. R.: Computer Movie Simulating Urban Growth in Detroit Region, Econ. Geogr., 46, 234-240, https://doi.org/10.2307/143141, 1970.

United States Army Coprs of Engineers (USACE): National Inventory of Dams (NID) [data set], https://nid.usace.army.mil, 2018.

Vörösmarty, C. J., Meybeck, M., Fekete, B., Sharma, K., Green, P., and Syvitski, J. P. M.: Anthropogenic sediment retention: major global impact from registered river impoundments, Glob. Planet Change, 39, 169-190, https://doi.org/10.1016/S0921-8181(03)00023-7, 2003.

Wada, Y., Reager, J. T., Chao, B. F., Wang, J., Lo, M. H., Song, C., Li, Y. W., and Gardner, A. S.: Recent changes in land water storage and its contribution to sea level variations, Surv. Geophys., 38, 131-152, https://doi.org/10.1007/s10712-016-9399-6, 2017.

Wang, J., Sheng, Y., and Wada, Y.: Little impact of the Three Gorges Dam on recent decadal lake decline across China's Yangtze Plain, Water Resour. Res., 53, 3854-3877, https://doi.org/10.1002/2016WR019817, 2017.

Wang, J., Walter, B.A., Yao, F., Song, C., Ding, M., Maroof, M.A.S., Zhu, J., Fan, C., McAlister, J.M., Sikder, M.S., Sheng, Y., Allen, G.H., Crétaux, J.-F., and Wada, Y., 2021. GeoDAR: Georeferenced global dams and reservoirs dataset for bridging attributes and geolocations. Earth System Science Data, in review.

Whittemore, A., Ross, M. R. V., Dolan, W., Langhorst, T., Yang, X., Pawar, S., Jorissen, M., Lawton, E., Januchowski-Hartley, S., and Pavelsky, T.: A participatory science approach to expanding instream infrastructure inventories, Earth's Future, 8, e2020EF001558, https://doi.org/10.1029/2020EF001558, 2020.

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., and Pavelsky, T. M.: MERIT Hydro: A high-resolution global hydrography map based on latest topography dataset, Water Resour. Res., 55, 5053-5073, https://doi.org/10.1029/2019WR024873, 2019.

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C., Kanae, S., and Bates, P. D.: A high-accuracy map of global terrain elevations, Geophys. Res. Lett., 44, 5844-5853, https://doi.org/10.1002/2017GL072874, 2017.

Yang, X., Pavelsky, T. M., Ross, M. R. V., Januchowski-Hartley, S. R., Dolan, W., Altenau, E. H., Belanger, M., Byron, D., Durand, M., Van Dusen, I., Galit, H., Jorissen, M., Langhorst, T., Lawton, E., Lynch, R., Mcquillan, K. A., Pawar, S., and Whittemore, A.: Mapping flow-obstructing structures on global rivers, Water Resour. Res., 58, e2021WR030386, https://doi.org/10.1029/2021WR030386, 2022.

Yao, F., Wang, J., Wang, C., and Cretaux, J. F.: Constructing long-term high-frequency time series of global lake and reservoir areas using Landsat imagery, Remote Sens. Environ., 232, 111210, https://doi.org/10.1016/j.rse.2019.111210, 2019.

Yassin, F., Razavi, S., Elshamy, M., Davison, B., Sapriza-Azuri, G., and Wheater, H.: Representation and improved parameterization of reservoir operation in hydrological and land-surface models, Hydrol. Earth Syst. Sci., 23, 3735-3764, https://doi.org/10.5194/hess-23-3735-2019, 2019.

Yigzaw, W., Li, H. Y., Demissie, Y., Hejazi, M. I., Leung, L. R., Voisin, N., and Payn, R.: A new global storage-area-depth data set for modeling reservoirs in land surface and earth system models, Water Resour. Res., 54, 10372-10386, https://doi.org/10.1029/2017WR022040, 2018.

Zarfl, C., Lumsdon, A. E., Berlekamp, J., Tydecks, L., and Tockner, K.: A global boom in hydropower dam construction, Aquat. Sci., 77, 161–170, https://doi.org/10.1007/s00027-014-0377-0, 2015.

Zhan, S., Song, C., Wang, J., Sheng, Y., and Quan, J.: A global assessment of terrestrial evapotranspiration increase due to surface water area change, Earth's Future, 7, 266-282, https://doi.org/10.1029/2018EF001066, 2019.

Zhang, S., Gao, H., and Naz, B. S.: Monitoring reservoir storage in South Asia from multisatellite remote sensing, Water Resour. Res., 50, 8927-8943, https://doi.org/10.1002/2014WR015829, 2014.

Zhang, W., Pan, H., Song, C., Ke, L., Wang, J., Ma, R., Deng, X., Liu, K., Zhu, J., and Wu, Q. H.: Identifying emerging reservoirs along regulated rivers using multi-source remote sensing observations, Remote Sens-Basel, 11, 25, https://doi.org/10.3390/rs11010025, 2019.

Zhao, G. and Gao, H.: Estimating reservoir evaporation losses for the United States: Fusing remote sensing and modeling approaches, Remote Sens. Environ., 226, 109-124, https://doi.org/10.1016/j.rse.2019.03.015, 2019a.

Zhao, G. and Gao, H.: Towards global hydrological drought monitoring using remotely sensed reservoir surface area, Geophys. Res. Lett., 46, 13027-13035, https://doi.org/10.1029/2019GL085345, 2019b.