Earth Syst. Sci. Data, 10, 765–785, 2018
https://doi.org/10.5194/essd-10-765-2018


Open Access Earth System Science Data

# The Global Streamflow Indices and Metadata Archive (GSIM) – Part 1: The production of a daily streamflow archive and metadata

**Hong Xuan Do[1], Lukas Gudmundsson[2], Michael Leonard[1], and Seth Westra[1]**

[1]School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, Australia
[2]ETH Zürich, Institute for Atmospheric and Climate Science, Zürich, Switzerland

**Correspondence:** Hong Xuan Do (hong.do@adelaide.edu.au)

Received: 7 September 2017 – Discussion started: 20 September 2017
Revised: 9 March 2018 – Accepted: 16 March 2018 – Published: 17 April 2018

**Abstract.** This is the first part of a two-paper series presenting the Global Streamflow Indices and Metadata archive (GSIM), a worldwide collection of metadata and indices derived from more than 35 000 daily streamflow time series. This paper focuses on the compilation of the daily streamflow time series based on 12 free-to-access streamflow databases (seven national databases and five international collections). It also describes the development of three metadata products (freely available at https://doi.pangaea.de/10.1594/PANGAEA.887477): (1) a GSIM catalogue collating basic metadata associated with each time series, (2) catchment boundaries for the contributing area of each gauge, and (3) catchment metadata extracted from 12 gridded global data products representing essential properties such as land cover type, soil type, and climate and topographic characteristics. The quality of the delineated catchment boundary is also made available and should be consulted in GSIM application. The second paper in the series then explores production and analysis of streamflow indices. Having collated an unprecedented number of stations and associated metadata, GSIM can be used to advance large-scale hydrological research and improve understanding of the global water cycle.

## 1 Introduction

Streamflow observations with global coverage are essential to make progress in the science of large-scale hydrology. For example, global datasets provide particular value when evaluating global hydrological models (Gudmundsson et al., 2012; Huang et al., 2016; Ward et al., 2013), producing runoff estimation data products (Fekete et al., 2002a, b; Gudmundsson and Seneviratne, 2015; Vörösmarty et al., 1989), investigating large-scale weather patterns and their relation to hydrological extremes (Wanders and Wada, 2015; Ward et al., 2014), and detecting changes in the global hydrological extremes over space and time (Do et al., 2017; Gudmundsson et al., 2017; Kundzewicz et al., 2012; Milly et al., 2002), amongst numerous other applications.

Despite the fundamental, widespread, and varied applications that streamflow observations support, there are many obstacles to the existence and utility of a large-scale streamflow archive. Firstly, there are threats to the quantity of data, such as political sensitivities (Nelson, 2009), cost recovery and strict access policies (Hannah et al., 2011), unavailability in an electronic format, consistency of data formats, limited documentation, missing metadata, and a lack of resources for database maintenance and updating. Secondly, there are difficulties associated with the quality of the data in many regions, such as poor spatial coverage, poor quality control, variable quality control between regions, inconsistent metadata, imprecise geographic coordinates of the site, changes in the density of stream gauges, and variable record lengths. Lastly, even in locations where there are abundant and high-quality streamflow observations, there can be questions over its utility in specific research such as climate sensitivity analysis due to the manifestation of human impacts – for example, urbanization, land-use changes, channelization, and upstream dams (Hannah et al., 2011).

To date, the Global Runoff Data Base (GRDB) maintained by the Global Runoff Data Centre has been the primary dataset used in large-scale hydrological studies, with more than 9000 stations available to the research community (GRDC, 2015). The Global Runoff Data Centre (GRDC) database operates under the auspices of the UN – World Meteorological Organization (WMO), and its database is supported on a voluntary basis so that the number of data submissions depends on contributions by national authorities. However, although numerous countries have databases of acceptable quality, data supply remains resource intensive and the GRDB remains sparse in some regions. For example, the latest catalogue of the GRDB database (version 5 December 2017) shows that out of 7238 daily time series, there are only 637 stations over South America and only 642 stations over Asia. Moreover, many stations in regions such as Asia and Russia have not been updated for many years and are missing otherwise available data at the end of their records.

The Global Streamflow Indices and Metadata (GSIM) project has been initiated in order to address the demand for a global streamflow database (Bierkens, 2015; Fekete et al., 2015; Hannah et al., 2011; Kundzewicz et al., 2013; Merz et al., 2012; Milly et al., 2015). The approach of this project is not to collect high-quality data from referenced hydrological networks, which have been conducted in other studies (Addor et al., 2017; Burn et al., 2012; Hannaford and Marsh, 2006; Hodgkins et al., 2017; Whitfield et al., 2012) to support research that requires assumptions regarding the minimum impact of human interference on streamflow, such as the investigation of climate change implication for changes in extreme events. Instead, the activities of the GSIM project have been to collate publicly available data, apply basic consistency to the formatting, and establish a standardized set of metadata. In so doing, GSIM intends to promote more widespread use of streamflow data, facilitate improved research outcomes through increased spatial coverage and gauge density, and tackle ongoing challenges for the hydrological community, for example, addressing fundamental issues of data quality, identifying additional data sources, lobbying for continuity of data networks, and developing a method for improved governance and maintenance of streamflow data at the global scale.

To maximize the value of the streamflow dataset for a wide range of applications, the GSIM project also seeks to provide information on catchment characteristics upstream of the streamflow gauging station. This necessitates a consistent approach to delineating the upstream catchment boundary for every gauge station, and this is achieved using data from a global digital elevation model (DEM). This is because, with the exception of the GRDB databases, catchment boundaries representing the direct drainage area of stations were unavailable. Filling in this missing element of metadata is important to facilitate further analysis of the streamflow observations with respect to a wide and ever-increasing variety of spatial datasets. Although there have been previous efforts in producing catchment boundaries for a smaller number of stations (Addor et al., 2017; Arsenault et al., 2016; Lehner, 2012; Schaake et al., 2006), similar work at this magnitude has not been undertaken. This task is complicated by a lack of precision in the supplied geographic coordinates of a given site; for example, when a catchment boundary is extracted, the corresponding calculated area may not match the reported area of the catchment and a procedure for checking minor shifts in the coordinates is needed to improve identification of the likely catchment boundary. The quality of the delineated catchment boundary is also made available to GSIM users and should be considered prior to using this data product and any accompanying information.

The availability of catchment boundaries for each gauge enables the association of environmental variables with each gauge by extracting them from corresponding global-scale gridded products. As part of the GSIM project, a number of global data products are provided as an additional dataset so that a user can readily filter the GSIM dataset according to specific interests, for example, by climate type, soil type, land-use type, irrigation area, and population density. Other potential applications of this auxiliary information might include a comparison to a database of dams for identifying upstream impacts; to remotely sensed estimates of forest cover or urban extent for determining land-use change; to population demographics for improving estimates of flood exposure; and to hydrological model outputs for evaluating model performance.

Finally, to facilitate benefits of this project to the broader community, indices characterizing water-balance aspects, hydrological extremes, and features of the seasonal cycle have been derived from the GSIM time series and will be made publicly available. To ensure standardized quality for the derived indices, a quality control procedure coupling the information provided by data providers and a data-driven approach was also applied.

This is the first paper of a two-part series detailing the production of GSIM and corresponding data products. This paper outlines the provenance of daily streamflow time series (Sect. 2), procedures for reformatting and combining the time series (Sect. 3), the development of metadata associated with each gauge (Sect. 4), an overall summary of the GSIM time series and metadata (Sect. 5), and data availability (Sect. 6). As the time-series database cannot be made available online due to varieties of terms and conditions from data providers, the second paper in this series (Gudmundsson et al., 2018) is dedicated to the production of streamflow time-series indices, including (1) checks for data quality, (2) the production of streamflow time-series indices, and (3) homogeneity assessment of the derived indices.

## 2 Daily streamflow data and where to find them

GSIM is a compilation of 12 databases that have either open-access or restricted-access policies, and that collectively represent a total of 35 002 stations. The spatial distribution and the number of stations available in each database are illustrated in Fig. 1 (continental-scale figures are also provided as a Supplement). A summary of the data sources is also provided in Table 1 and detailed information on each database is elaborated upon in the following sections. The list of databases identified as part of GSIM is not exhaustive of all possible data sources, only of those that were known to the authors and readily accessible within the project time frame. Where additional data are available in a convenient format, it may be possible to further augment GSIM in the future.

The various data sources were classified as either a "research database" or a "national database". The reasons for this classification are further outlined in Sect. 3, but relate to issues when merging databases and removing duplicate gauges. The data sources include the following.

1. Research databases: databases with daily streamflow data that have been compiled on an ad hoc basis from a variety of original sources by research organizations. This category includes five different databases: the Global Runoff Data Base (GRDB); the European Water Archive (EWA); the China Hydrological Data Project (CHDP) data archive; the GEWEX Asian Monsoon Experiment – Tropics (GAME) data archive; and the Regional Hydrographic Data Network for the Arctic Region (ARCTICNET) data archive.

2. National databases: databases with daily streamflow data made publicly available by national water authorities as part of water-related regulations. This category includes seven databases: the USGS Water Data for USA database (USGS); Canada's National water data archive (HYDAT); Japan's Ministry of Land and Infrastructure database Water Information System (MLIT); Spain's digital hydrological yearbook database (Anuario de aforos digital 2010–2011, AFD); Australia's Bureau of Meteorological Water Data Online database (BOM); India's Water Resources Information System database (WRIS); and Brazil's National Water Agency database (ANA).

### 2.1 The Global Runoff Data Base (GRDB)

The daily streamflow dataset received from the GRDC (6313 stations with more than 10 years on record; see also Gudmundsson and Seneviratne, 2016) is referred to as the GRDB in this project. To date, the GRDB has been the largest and most extensively used dataset for streamflow analysis at regional and global scales. It was thus considered as the starting point and "base" for the GSIM project. Indeed, it was awareness of data not available from the GRDB that prompted the initial search for additional sources of data to complement the database.

The GRDC was initiated in 1988 by the WMO and is now maintained at the German Federal Institute of Hydrology in Koblenz. The GRDC provides free and unrestricted access to all hydrological data and products, although the data policy indicates that requests for data must reach the GRDC in written form to ensure data users do not redistribute the time series. More detail about the GRDC data policy, and the procedure for obtaining its time series, are outlined at http://www.bafg.de/GRDC/EN/01_GRDC/12_plcy/data_policy_node.html (last access: 23 June 2017).

### 2.2 The European Water Archive

The European Water Archive (referred to as the EWA in this paper) is one of the most comprehensive streamflow time-series archives in Europe, with more than 3000 river gauging stations distributed across 29 countries. This archive is also currently held by the GRDC and available under the GRDC data policy (http://www.bafg.de/GRDC/EN/04_spcldtbss/42_EWA/ewa_node.html, last access: 3 January 2018). The EWA stations used in this paper were selected using the same criteria as Gudmundsson and Seneviratne (2016), with a total of 3731 daily records.

### 2.3 The China Hydrology Data Project

The China Hydrology Data Project (CHDP) aims to digitize an arrangement of hydrological measurements taken at Chinese stations. These measurements (including daily discharge) were originally only available in book form (Henck et al., 2010). The original data were collected by the Chinese Hydrology Bureau and published in annual yearbooks. At the time GSIM began, discharge data were only available for the Yunnan-Tibet International Rivers, which corresponded to 163 stations until 1987. This project has been terminated since the 2000s and thus no further update is available. The data and metadata were obtained directly from the author of the project. Detailed information can be viewed at http://www.oberlin.edu/faculty/aschmidt/chdp/index.html (last access: 23 June 2017).

### 2.4 The GEWEX Asian Monsoon Experiment – Tropics project

The GEWEX Asian Monsoon Experiment – Tropics project (GAME) was initiated in 1996 to monitor several hydroclimatological variables over the humid temperate area in South-East Asia. As one of several important activities in this project, many hydrological observation datasets were collected, including streamflow data. Available streamflow data were provided by the Royal Irrigation Department of Thailand, and comprised 129 time series spanning a period from 1980 to 2000. Daily discharge data and associ-

**Table 1.** Basic information of daily streamflow databases included in the GSIM project.

| Database (referred name) | Database category | Spatial coverage | Data access information |
|---|---|---|---|
| Global Runoff Data Base (GRDB) | Research database | Global | www.bafg.de/GRDC/ (last access: 23 June 2017) Archived database can be obtained via written request to the Global Runoff Data Centre. This database is updated when new data are submitted by national suppliers. |
| European Flow Regimes from International Experimental and Network Data (EWA) | Research database | European | http://ne-friend.bafg.de/servlet/is/7413/ (last access: 23 June 2017) Data can be obtained via written request to the Global Runoff Data Centre. This database has been frozen since October 2014 and is being integrated into the GRDB database. |
| A Regional, Electronic, Hydrographic Data Network for Russia (ARCTICNET) | Research database | Russia | http://www.russia-arcticnet.sr.unh.edu/ (last access: 23 June 2017) Archived and closed historic database. Part of this data archive has been included in the databases of the Global Runoff Data Centre and updated based on data deliveries. |
| China Hydrology Database Project (CHDP) | Research database | China | http://www.oberlin.edu/faculty/aschmidt (last access: 23 June 2017) Archived and closed historic database can be obtained via written request to the author of the database. |
| GEOSS ana MAHASRI Experiment in Tropics (GAME) | Research database | Thailand | http://hydro.iis.u-tokyo.ac.jp/GAME-T/GAIN-T/routine/rid-river/disc_d.html (last access: 23 June 2017) Archived and closed historic database |
| US National Water Information System (USGS) | National database | USA | http://waterdata.usgs.gov/nwis (last access: 23 June 2017) Individual time series can be downloaded from the data portal (updated regularly). |
| Canada National Water Data Archive (HYDAT) | National database | Canada | https://ec.gc.ca/rhc-wsc/ (last access: 23 June 2017) Archived database. The archive is updated quarterly by the data authority. |
| Brazil National Water Agency (ANA) | National database | Brazil | http://hidroweb.ana.gov.br/ (last access: 23 June 2017) Individual time series can be downloaded from the data portal (updated regularly). |
| Japan Water Information System (MLIT) | National database | Japan | http://www1.river.go.jp/ (last access: 23 June 2017) Individual time series can be downloaded from the data portal (updated regularly). |
| Anuario de aforos digital 2010–2011 (AFD) | National database | Spain | http://ceh-flumen64.cedex.es/anuarioaforos (last access: 23 June 2017) Archived database, DVD available from Spanish authorities (updated annually) |
| Australia Water Data Online (BOM) | National database | Australia | http://www.bom.gov.au/waterdata/ (last access: 23 June 2017) Individual time series can be downloaded from the data portal (updated regularly). |
| Water Resources Information System of India (I-WRIS) | National database | India | http://www.india-wris.nrsc.gov.in/wris.html (last access: 23 June 2017) Individual time series can be downloaded from the data portal (updated regularly). |

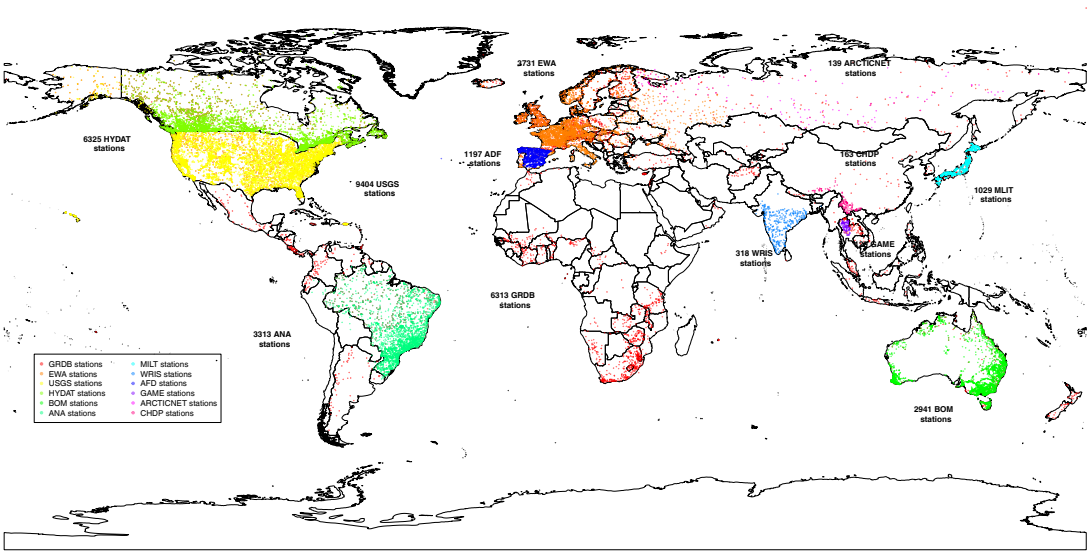


**Figure 1.** The distribution of stations from original data sources.

ated metadata were archived and can be accessed online at http://hydro.iis.u-tokyo.ac.jp/GAME-T/GAIN-T/routine/rid-river/index.html (last access: 23 June 2017).

### 2.5 The ARCTICNET project

A regional hydrometeorological data network for the pan-Arctic Region project is a regional database that can be accessed via the Internet and is referred to as ARCTICNET in this paper. The database is designed to support hydrological sciences and water resource assessments over this region with the goal of estimating the contemporary water and constituent balances for the pan-Arctic drainage system. ARCTICNET is a static dataset and some time series have been included in the databases of the GRDC and updated based on data deliveries. Although most data provided in the data portal are at monthly resolution, there are 139 high-quality daily streamflow time series across Russia that are also available which have not been fully integrated into GRDB. Although ARCTICNET's future status is likely to be a part of the GRDB, these stations have still been considered in GSIM production and are referred to as the ARCTICNET database in this paper. These time series, along with their metadata, were archived and can be downloaded at http://www.r-arcticnet.sr.unh.edu/v4.0/index.html (last access: 23 June 2017).

### 2.6 The USGS database

The USGS National Data Services for the US provide access to water resources data collected at approximately 1.5 million sites in all 50 states of the USA, also including the District of Columbia, Puerto Rico, the Virgin Islands, Guam, American Samoa, and the Commonwealth of the Northern Mariana Islands. All time series and associated metadata can be queried from the data portal http://waterdata.usgs.gov/nwis (last access: 23 June 2017). To ensure the queried data have sufficient geographic metadata (critical for the present project), the stations listed in the Geospatial Attributes of Gages for Evaluating Streamflow, version II (GAGES II) database were used (Falcone, 2011). The time series from 9404 stream gauges obtained from the USGS data portal are referred to as the USGS database in this paper.

### 2.7 The HYDAT database

Canada's National Water Data Archive (HYDAT) is a database containing daily observed hydrometric data from publicly funded gauges in Canada. Also available in the HYDAT database are metadata about the hydrometric stations, such as latitude and longitude, catchment area, record length, as well as information regarding flow conditions (current status, regulated or natural regime). The database is updated four times per year and currently contains data for 6325 streamflow stations across Canada. The raw data, as well as an extractor executable, are publicly available from Environment Canada's website at https://ec.gc.ca/rhc-wsc/default.asp?lang=En&n=9018B5EC-1 (last access: 23 June 2017).

### 2.8 The ANA database

The HIDROWEB data portal was organized by the Brazilian National Water Agency (ANA). It provides a database with all the information collected by Brazil's hydrometeorological network. Streamflow data and associated metadata were made publicly available by Brazil's national water regulations, and have been used extensively to monitor critical

events, such as floods and droughts. Individual time series and their associated metadata can be viewed or downloaded at http://hidroweb.ana.gov.br (last access: 23 June 2017). The 3313 stations downloaded from this website are referred to as the ANA in this paper.

### 2.9 The AFD database

Spanish streamflow data were retrieved from the digital hydrological yearbook (Anuario de aforos digital 2010–2011, AFD), which provides observations until 2013–2014 and is freely accessible online at http://ceh-flumen64.cedex.es/anuarioaforos/default.asp (last access: 23 June 2017). For the GSIM, we used the time series that was used to develop the E-RUN dataset (Gudmundsson and Seneviratne, 2016). The original DVD containing the full database was obtained directly from the Spanish authorities via a written form request. This collection contains streamflow data from 1197 gauging stations, and is referred to as ADF in this paper.

### 2.10 The MLIT database

In Japan, the Ministry of Land, Infrastructure, Transport and Tourism is responsible for organizing hydrological data. All records are disseminated at http://www1.river.go.jp/ (last access: 23 June 2017). As of 2010, the database kept records of all river stations (at both discharge and gauge level). The composition of the 15-digit station IDs is outlined in the file http://www1.river.go.jp/kitei_sosoku.pdf (PDF), and can be used to query and download time series, along with its metadata. As the whole database is recorded in Japanese, the translateR package (Lucas and Tingley, 2016) was used to translate the metadata into English. The time series downloaded from the Japanese water data portal (1029 stations in total) is referred to as MLIT in this paper.

### 2.11 The BOM database

As part of the water reform programme established in Australia, Water Data Online was created to provide free access to nationally consistent, current and historical water information. It can be accessed at http://www.bom.gov.au/waterdata (last access: 23 June 2017). Water Data Online also contains historical data from some stations that are no longer operational. Users can view or download individual streamflow time series from the data portal, along with standardized data and reports. The time series measured at 2941 stations obtained from Water Data Online is referred to as the BOM database in this project.

### 2.12 The WRIS database

The Generation of Database and Implementation of Web Enabled Water Resources Information System in the Country project (India-WRIS WebGIS) was initiated as a joint venture of the Indian Central Water Commission (CWC) and the Indian Space Research Organization (ISRO). Unclassified data can be accessed online and free of charge at http://www.india-wris.nrsc.gov.in/wris.html (last access: 23 June 2017), while the metadata are documented at http://www.cwc.nic.in/main/downloads/HydrologicalnetworkdetailsofCWC.pdf (last access: 23 June 2017). All 318 stations were downloaded from the website. They are referred to as the WRIS database in this paper.

The production of time series and metadata for GSIM comprises several stages due to the range of data formats and significant variation in the quality of metadata across data sources. To ensure GSIM is presented in a transparent manner, the following sections outline procedures that are used to collate the time series across (Sect. 3), and to produce the metadata (Sect. 4).

## 3 Procedure for combining databases

Several of the identified data sources share common spatial domains, where typically the research databases may contain a subset of gauges from the national databases. It is therefore important to correctly identify duplicate time series when merging the databases. To maximize the quality of combined time series and minimize the requirement to combine time series, this task is conducted following three sequential steps: Step 1 – pre-processing the data to a common structure; Step 2 – replacing all GRDB stations in countries that have a national database; and Step 3 – identifying remaining duplicates. From the 35 002 gauges, 3197 (2958 and 239 gauges from the GRDB and EWA databases, respectively) were replaced by national databases in Step 2, and 846 cases of "very likely identical" stations were identified and removed in Step 3, leaving 30 959 "duplication-free" time series available in the GSIM.

### 3.1 Pre-processing the time series into a singular data structure

One of the major challenges in producing consistent streamflow indices is that data from different sources have different structures and storage formats. For example, the MLIT database divides streamflow records at one location into separate text files, and each file contains streamflow measurements for 1 year. In comparison, the HYDAT archive includes streamflow measurements from all available stations in a single matrix.

To address the varying standards of data management, the first step in combining the databases was to reformat all the streamflow records to ensure that each time series is kept in a consistent format. Using the GRDB as a guide, it was decided to store all data for a given site in a single text file with three columns: (a) date of measurement, (b) value of measurement, and (c) original quality flags (if available), and with basic metadata (station name, ID, etc.) stored in the header of the

file. All additionally derived metadata (i.e. from global gridded products) are stored in the station catalogue. The streamflow measurements were also converted into consistent units (cubic metres per second).

Metadata that have special characters in foreign language sources were also pre-processed into the ASCII encoding system. For river names and station names that are recorded in Spanish (ADF) or Portuguese (ANA), the special characters were replaced by plain alphabetic characters using the core function iconv() of the R programming language. For river names and station names that are recorded in Japanese characters (MLIT), R package "translateR" (Lucas and Tingley, 2016) was used with the Google Translate API for this task. Although there are some limitations related to this toolset (e.g. some Japanese characters remaining untranslated and requiring manual translation; inconsistency in the translated results using the same original Japanese characters), this option was chosen to enable an automated and expedient translation. As a result, any text-related metadata associated with Japanese stations should be treated with care.

### 3.2 Replace the GRDB stations with national databases, if applicable

The streamflow records hosted by the GRDC (the GRDB and EWA databases) are themselves originally provided by national water agencies, and have been undergone quality control procedures by the GRDC. In cases that the supplied data contain errors, the GRDC informs data suppliers to improve the quality of their database. In term of data availability, time series downloaded directly from the national data portal usually represents the latest version of streamflow observation, and thus it seemed appropriate to replace stations hosted by the GRDC for countries where an equivalent national database was available. While this approach is efficient, there is a potential downside of removing GRDB stations that were not otherwise present in the national data depositories, perhaps due to differences in maintenance of the databases. Nonetheless, the number of stations available in the GRDB and EWA databases is much lower than that available in national databases for all countries (see Table 2). As a result of this step, 2958 stations located in seven countries (Australia, Brazil, Canada, India, Japan, Spain, and the United States) were removed from the GRDB collection. In addition, 239 stations located in Spain were removed from the EWA archive.

### 3.3 Identify and remove duplicates in research databases

The method of de-duplicating time series involves identification of duplicates where two data sources have overlapping coverage and potential merging of two records at a duplicated site to create a unified record. The de-duplication step was generally undertaken between the GRDB and a "paired" dataset (e.g. GRDB and GAME). The only exceptions to this step are for GRDB, EWA, and ARCTICNET, as these three datasets share Russia as a common spatial domain.

**Table 2.** Number of stations in countries where national databases are available.

| Country | Database | | |
|---|---|---|---|
| | EWA | GRDB | National |
| Australia | – | 358 | 2941 (BOM) |
| Brazil | – | 439 | 3313 (ANA) |
| Canada | – | 1029 | 6325 (HYDAT) |
| India | – | 0 | 318 (WRIS) |
| Japan | – | 151 | 1029 (MLIT) |
| Spain | 239 | 0 | 1197 (ADF) |
| United States | – | 981 | 9404 (USGS) |

The techniques adopted for combining research databases were based on the de-duplication procedures developed in Gudmundsson and Seneviratne (2016), which consists of three sequential steps.

1. Identification of "duplication candidates" using metadata similarity. This step aims to identify time series with a high level of similarity in metadata (either within one database or across different databases). We used three similarity metrics to identify potential time series: (1) Jaro–Winkler distances, a metric representing the alphanumeric similarity of strings (Christen, 2012), applied to river names of two records; (2) Jaro–Winkler distances between station names of two records; and (3) geographical proximity estimated from geographical coordinates between two records. These metrics were normalized to have the same range between 0 and 1, where a value of 0 indicates identical metadata (e.g. the same geographic coordinates). This similarity analysis was run for each pair in the pool of stations, and any pair with an average value below 0.25 was identified as candidate duplicate records.

2. Classifications of duplication candidates using time-series similarity. This step aims to decide whether a specific pair of duplication candidates is likely to be identical. The overlapping period and correlation coefficient were used as criteria for making a decision. Firstly, all duplication candidates that do not share any overlap in their period of record are kept in the final GSIM collection, as they can represent separate time series even if they measured discharge at the same geographical location (e.g. due to reconstruction of the gauging station). Secondly, any time series with a correlation coefficient ($R^2$) lower than 0.90 was automatically identified as "very likely different" (26 pairs), whereas $R^2 > 0.99$ indicates "very likely identical" time series (786 pairs). Finally, candidates with $0.90 \leq R^2 \leq 0.99$

**Table 3.** Basic metadata available from data sources.

| Database | Station ID | Station name | River name | Geographical coordinates | Station elevation | Drainage area | Catchment boundary |
|---|---|---|---|---|---|---|---|
| GRDB | X | X | X | X | X | X | X |
| EWA | X | X | X | X | X | X | X |
| CHDP | X | X | X | X | – | X | – |
| GAME | X | X | X | X | X | X | – |
| ARCTICNET | X | X | X | X | X | X | – |
| USGS | X | X | – | X | X | X | – |
| HYDAT | X | X | – | X | – | X | – |
| ANA | X | E | E | X | X | X | – |
| ADF | X | E | E | X | X | X | – |
| MLIT | X | E | E | X | – | – | – |
| BOM | X | X | – | X | – | – | – |
| WRIS | X | X | X | X | X | X | – |

X: metadata available; –: metadata are unavailable; E: metadata are not available in English.

(65 pairs) were visually inspected and manually classified as "very likely identical" (60 pairs) or "very likely different" (five pairs). All time series in the "very likely different" category were retained while stations of the "very likely identical" category were processed using the de-duplication procedure (see below).

3. De-duplication of identical time series: regardless of whether identical time series come from either the same database or from different databases, records with the greater number of data points in the streamflow time series were kept while the other(s) were discarded. Although this approach has the downside of truncating the length of useful records, the number of time series that could be influenced by this approach is relatively low (846 time series, corresponding to 2.8 % of the total number of available streamflow records).

A visual example of the de-duplication procedure is provided in Fig. 2. The left panel demonstrates a case of "very likely identical" stations, when station number 2964035 in the GRDB database was identified as an identical gauge to W.16 in the GAME archive, based on the similarities between the provided metadata and correlation coefficient. The time series representing station "GAME_W.16" was kept in the final collection, while time series "GRDB_2964035" was removed. The right panel in Fig. 2 demonstrates a case of a "duplication candidate" with correlation coefficient of 0.92 (time series "GRDB_6123645" and "EWA_9110028"). These time series were visually inspected, assigned a "very likely different" label, and both time series were kept in the final collection.

## 4 Production of the GSIM metadata

Providing a consistent set of metadata for each site has been a significant undertaking for GSIM. This section outlines three main stages to developing the GSIM metadata: (1) consolidating all available basic metadata; (2) consistently delineating catchment boundaries for each site; and (3) developing a supplementary set of catchment-scale metadata based on the delineated boundaries.

### 4.1 Consolidating basic metadata from available sources

Following the GRDB format, each time series was accompanied by basic metadata, including

1. station ID,
2. station name,
3. river name of gauging location,
4. geographical coordinates of station,
5. elevation of station,
6. drainage area, and
7. catchment boundary from original data sources.

These data are useful for filtering stations according to specific criteria and analysis objectives. Moreover, the availability of a catchment boundary for the gauge enables additional catchment-scale metadata to be derived as necessary. However, not all of these basic metadata were available for all data sources. For example, the catchment boundary was only available for parts of the GRDB and EWA stations, the drainage area was unavailable in the BOM and MLIT databases, and though several data sources included river names in station names (BOM, HYDAT, USGS), these metadata were unavailable in English for other sources (MLIT, ANA, ADF). Table 3 further outlines the availability of basic metadata for each source.

The method for consolidating basic metadata for each station follows three steps.

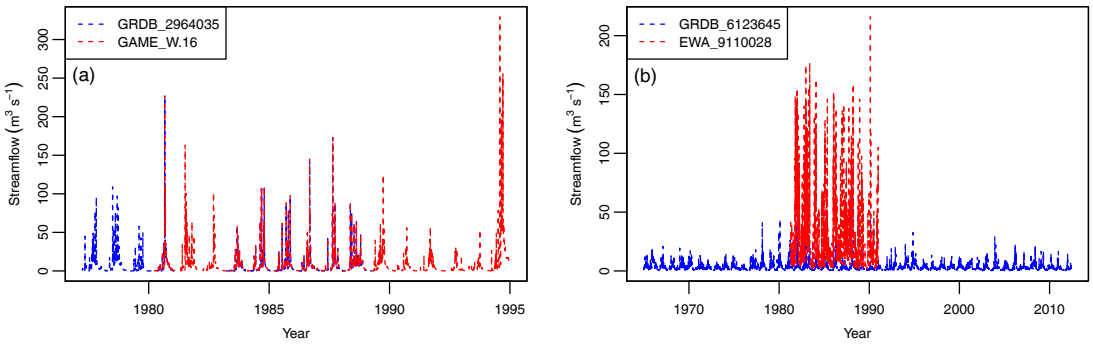


**Figure 2.** Examples of visually inspected duplication-candidate time series. **(a)** Two stations that were labelled "very likely identical" stations. **(b)** Two stations that were labelled "very likely different" stations.

### Step 1. Transfer and review metadata available from original sources

The transfer of all existing metadata required a range of simple consistency checks and conforming rules, including the following.

1. Reviewing the geographical coordinates of all stations. Stations with unreasonable locations (e.g. located in the middle of the North Atlantic Ocean without any land mass, identified from Google Earth) were marked to be excluded from the subsequent delineation procedure (24 stations).

2. Separating the river name from the station name. Several sources use a consistent format for the station name consisting of two parts: the name of the station followed by the name of the water body. This pattern used a formula with "linking words" such as "at", "upstream" and "downstream". Taking station "BOM_406219" with original station name "Campaspe River at Lake Eppalock (Head Gauge)" as an example, the position of linking word "at" was identified and used to extract "river" metadata (Campaspe River) from the full station name.

3. Retaining the metadata of duplicated time series with the most data points in contrast to the other time series being removed. While this step may mistakenly remove some information, it is expedient and reflects the typical result of de-duplicated records that longer time series were kept while the shorter time series were removed.

### Step 2. Generate "database-merging" information

This step documents a summary of efforts taken in creating a consistent set of GSIM metadata, and allows a user to check steps that were taken or to identify better procedures using alternative time series or metadata obtained from original sources. There are 12 fields documented for this purpose:

1. an indication of whether the time-series de-duplication procedure was used (one field),

2. which database and station were kept to construct the GSIM time series (two fields),

3. which station was removed and the corresponding database (three fields),

4. the value of metrics that represent similarities in the time-series metadata (five fields), and

5. the number of overlapping days, if applicable (one field).

### Step 3. Generate information about data availability

The last step in compiling basic metadata for GSIM was to generate metrics that represent data availability for each GSIM time series, including the temporal coverage (i.e. the first and final years), the number of available daily observations, the number of missing data points, and the proportion of missing data points.

## 4.2 Catchment delineation procedure

With the ever-increasing availability of remote-sensing and modelled data products at global and continental scales, the provision of catchment boundaries is an important mechanism for extending the utility of GSIM. Although catchment boundaries can be generated easily using standard delineation algorithms in GIS packages, it requires a global coverage DEM dataset and reliable location to represent the outlet of each drainage area, which were unfortunately not readily available for GSIM project. This section describes the DEM products, and the algorithm to identify the "best outlet location" associated with each station that has been used in GSIM project.

The main DEM product used for GSIM was HydroSHEDS (http://hydrosheds.org, last access: 23 June 2017), which is

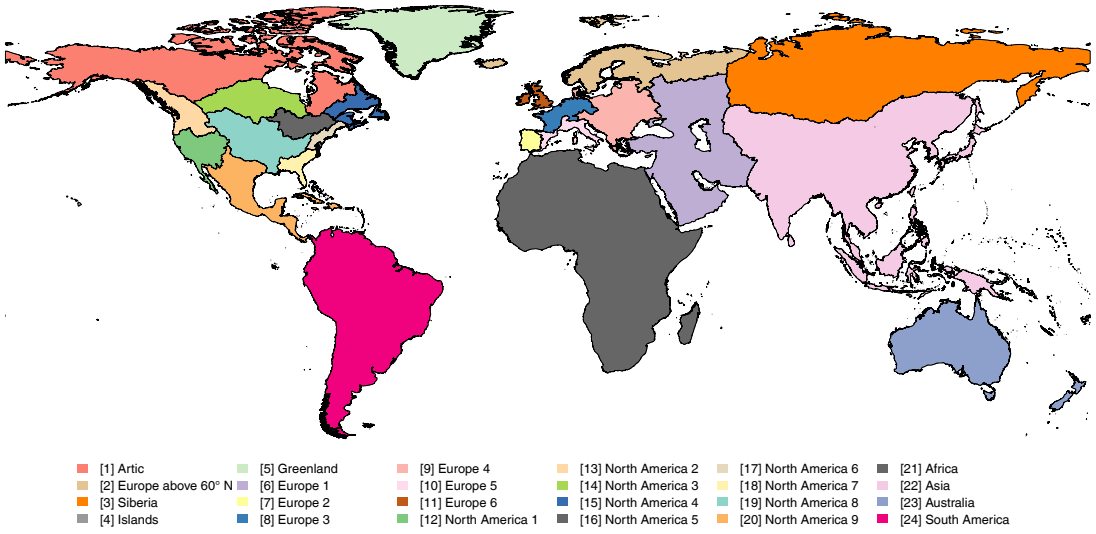


**Figure 3.** GSIM regions for catchment delineation and metadata extraction procedures.

available at 15 arcsec resolutions (Lehner et al., 2006), and has been used extensively in large-scale hydrological studies (Do et al., 2017; Lehner and Grill, 2013; Lehner et al., 2008; Wood et al., 2011). To address a limitation in the coverage of HydroSHEDS (no information in regions above 60° N, and some islands), the Viewfinder Panoramas elevation product at 15 arcsec resolutions was used (http://viewfinderpanoramas.org, last access: 25 June 2017) for those locations. This dataset has been used in several studies as an alternative DEM product to overcome similar data coverage issues (Barr and Clark, 2012; Fredin et al., 2012; Sil and Sitharam, 2016; Yamazaki et al., 2015). As there were more than 30 000 stations needing to be delineated, the HydroBASINS dataset was used, dividing the world into 24 regions, so that the task of delineation could be performed in parallel. The regions are shown in Fig. 3 and are generally independent in terms of drainage areas (Lehner and Grill, 2013). North America and Europe were specifically broken into more regions to address their relatively higher density of gauges. To maintain consistency when delineating boundaries, only one DEM product was used per GSIM region. As the quality of the Viewfinder Panoramas is not as clearly documented as for HydroSHEDS, its use was kept to a minimum. This resulted in five regions using Viewfinder DEM and 19 regions using HydroSHEDS (see Table 4).

Other challenges in the catchment delineation procedure are possible errors in the geographical coordinates representing the catchment outlet, such as typos in reported coordinates (e.g. 13.47° N instead of 14.47° N) or swapped order of the coordinate digits (e.g. 103.45° E instead of 103.54° E). These errors can lead to unreliable results of the delineation procedure, and so an algorithm to identify a location that represents catchment outlets well was also applied. This is described below.

### Case 1. Reported station coordinates adopted as the outlet

If there was no information about a drainage area in the station metadata, the geographical coordinates of the station available from the data source were used as the outlet of the delineation process. There are automated techniques for repositioning outlets, such as choosing cells with the greatest flow accumulation within a search distance (Snap Pour Point ArcGIS tool), or finding the nearest cell possessing a flow-accumulation value above a specified threshold (Lindsay et al., 2008). Nonetheless, without information on the catchment area, it is impossible to assess the quality of the delineated catchment. Even if a repositioning technique were adopted, delineated catchment boundaries should be used with caution in this case, and therefore the original geographical coordinates was used to represent "best outlet location".

### Case 2. Application of an automated repositioning algorithm

For stations with available information on catchment area, the automated repositioning procedure documented in GRDC report number 41 (Lehner, 2012) was used with some minor adjustments, and is summarized below.

1. The catchment area was estimated using the flow-accumulation dataset derived from the DEM products. This calculation was repeated for all pixels of the HydroSHEDS/Viewfinder gridded river network within a search radius of 5 km from the geographical coordinates of a specific station.

2. The estimated area values were compared with the reported area in the original metadata. All pixels were coded with the absolute value of their area differences

**Table 4.** DEM products used for each GSIM region.

| Region | Description | DEM product |
|---|---|---|
| Arctic (region 1) | Represents the distant part of North America (including Alaska, most parts of Canada, and the eastern part of Autonomous Province, Russia) | Viewfinder DEM 15s |
| Europe above 60° N (region 2) | Represents countries located above 60° N (e.g. Sweden, Denmark, Norway, part of Germany, part of Russia) | Viewfinder DEM 15s |
| Siberia (region 3) | Represents areas above the 60° N part of Asia | Viewfinder DEM 15s |
| Islands (region 4) | Represents some islands across the Pacific Ocean (e.g. Honolulu, US) and Atlantic Ocean | Viewfinder DEM 15s |
| Greenland (region 5) | Represents land mass of Greenland | Viewfinder DEM 15s |
| Europe 1 to Europe 6 (six regions, from region 6 to region 11) | Represent most European countries (below 60° N) | HydroSHEDS DEM 15s |
| North America 1 to North America 9 (nine regions, from region 12 to region 20) | Represent US (except Alaska) and the southern part of Canada (below 60° N). It also includes central America for simplicity in processing catchment boundaries. | HydroSHEDS DEM 15s |
| Africa (region 21) | Represents Africa region | HydroSHEDS DEM 15s |
| Asia (region 22) | Represents Asia region (part of Kazakhstan, China, Mongolia, and Russia) | HydroSHEDS DEM 15s |
| Australia (region 23 | Represents Australia, New Zealand, and some Pacific islands | HydroSHEDS DEM 15s |
| South America (region 24) | Represents South America | HydroSHEDS DEM 15s |

(in %, with reported area in the metadata used as a reference). Pixels with area differences of more than 50 % were excluded. This procedure provided an area-based ranking scheme (RA) ranging from 0 to 50, where 0 indicates perfect agreement in catchment areas.

3. The distance to the original location of the station (geographical coordinates reported in the original metadata) was calculated for each pixel and normalized to reach 50 at the maximum distance of 5 km. This procedure provided a distance-based ranking scheme (RD) ranging from 0 to 50, where 0 indicates perfect agreement in station locations.

4. The final ranking scheme ($R$) was calculated as a combination of RA and RD, where distance rank was weighted twice as high ($R = \mathrm{RA} + 2\mathrm{RD}$) to penalize pixels that were further away from the original location.

5. The outlet was automatically relocated to the position of the pixel showing the lowest ranking value, and geographical coordinates of the pixel centroid were defined as the "best" outlet for this specific catchment.

6. In the original technical document (Lehner, 2012), a manual procedure was adopted for stations with differences in area above 50 (i.e. the search algorithm cannot find any pixel with an area difference less than 50 % within the 5 km search radius), or for stations that had no reported area in the data catalogue. This manual inspection process was infeasible given the scope of the GSIM project, having over 30 000 catchments being delineated and where river names were not available (or potentially inaccurately translated) for many stations.

A Python script was developed to automatically call the "best outlet location" algorithm and the catchment delineation toolset available in ArcGIS software (Jenson and Domingue, 1988) for each gauge using the chosen DEM data product. The delineated catchment boundary for each station was assigned a quality flag according to the discrepancy between reported drainage area and delineated catchment boundary area. There are four quality categories associated with the catchment boundary:

1. "High" quality: Area difference less than 5 %

2. "Medium" quality: Area difference from 5 % to less than 10 %

3. "Low" quality: Area difference from 10 % to less than 50 %

4. "Caution" quality: Area difference greater than or equal to 50 %, or the reported catchment area was not available in the GSIM catalogue.

Figure 4 demonstrates an example where the repositioning algorithm was used. Here the "best outlet location" was determined to be 4.8611 km away from the original location, which is defined by the reported geographical coordinates in the metadata (for station AR_0000007). The reported area in the metadata is 340 $km^2$, while the area of the delineated catchment boundary using the original coordinates was only 0.8 $km^2$, which is significantly lower than the correct number. On the other hand, the delineated catchment boundary using the "best outlet location" has an area of 363 $km^2$, indicating a better estimation of the upstream catchment boundary for this particular station.

### 4.3 Extraction of catchment-scale metadata

An important aspect of large-scale hydrology is the ability to exploit gridded datasets at the global scale (Bierkens, 2015; Bierkens et al., 2015; Gudmundsson and Seneviratne, 2015; Seneviratne et al., 2012; Ward et al., 2015). Having developed catchment boundaries for each GSIM station enabled a supplementary set of catchment-scale metadata to be derived with relative ease. A key feature is that the catchment boundaries and the subsequent metadata relates to the upstream contributing area that influences a gauge, rather than to the catchment (or arbitrarily defined sub-catchment) that contains the gauge and therefore includes a non-influencing downstream region.

In developing the catchment-scale metadata, a standard set of variables have been identified with a view to supporting a range of applications such as filtering stations according to characteristic features, performing analyses of streamflow according to explanatory features of a catchment, or classifying stations according to the (in)significance of human impact. As summarized in Table 5, a total of 12 global data products were used to derive 19 elements of catchment-scale metadata. These products were chosen to represent five main categories of catchment characteristics: (1) topography, (2) human impact, (3) climate type, (4) vegetation type, and (5) soil profile. Because the global data products have varying resolution and structure, the following method was used to derive the catchment-scale metadata.

1. Delineated catchment boundaries associated with each stream gauge were used to mask the subset of pixels from the resampled dataset.

2. If more than 30 % of the catchment area was not covered by a specific global data product, a "No data" code was given.

3. Metadata representing the characteristics of the upstream catchment for each streamflow gauge were calculated from the gridded data masked in step (1). There were three types of metrics calculated during this step.

   a. A single value. Used only for the elevation at the geographical coordinates of the gauge (i.e. the catchment outlet), number of large dams located within the catchment boundary, and total volume of corresponding reservoir.

   b. Average, min, max, and quartile values. Used for continuously varying data such as a slope or topography index. These metrics allow an idea of central tendency as well as spread of extracted data within each catchment boundary.

   c. Percentages of different classes of catchment characteristics. Used for categorical data. For example, there are 16 classes in the global lithology dataset, and the co-presence of more than one type of lithology occurs very often across all catchments. The percentages of each lithology class were therefore calculated and recorded for all available catchments. To make the results presentable in a final catchment-scale metadata matrix, an aggregated metric was calculated to indicate that there is a dominant class within the catchment boundary (i.e. more than 50 % of all available pixels). If there is no dominant class within the catchment boundary, a "No dominant class" string is provided.

## 5 Overview of the GSIM archive

This section summarizes the GSIM archive, including the availability of time series combined from 12 original data sources, the associated data products, and documentation outlining data quality (Sect. 5.1). The whole time-series database cannot be made available online due to data policies from a number of original data sources, some of which apply very strict terms and conditions regarding the redistribution of streamflow time series. To address this limitation and maintain the usefulness of GSIM to the research community, three metadata products have been developed and the availability of these data products is further discussed in Sect. 5.2.

### 5.1 Time-series availability

From the 35 002 time-series records obtained from 12 different sources, the final GSIM time-series archive holds a total of 30 959 unique stations, of which 30 935 stations have associated catchment shapefiles and catchment-scale metadata (24 stations were removed from this process due to suspect geographical locations). Most data sources are still active and being updated by the data authorities. GSIM, however, also included 425 "static" time series (from the ARCTICNET, GAME, and CHDP databases) that have been frozen since the early 2000s as these stations have improved the gauge

**Table 5.** Global data products used in GSIM and derived catchment-scale metadata.

| Variables | Data sources | Spatial resolution | Reference period | Extracted metadata |
|---|---|---|---|---|
| Elevation | HydroSHEDS http://hydrosheds.org/ (last access: 23 June 2017) ViewFinder http://viewfinderpanoramas.org/ (last access: 23 June 2017) | 15 arcsec × 15 arcsec | – | (1) Gauge elevation (2a–f) Average, minimum, maximum, first quartile, second quartile, and third quartile values of catchment elevation |
| Slope | Derived from HydroSHEDS and ViewFinder DEM by authors | 15 arcsec × 15 arcsec | – | (3a–f) Average, minimum, maximum, first quartile, second quartile, and third quartile values of catchment slope |
| Topographic index | High-resolution global topographic index values (Marthews et al., 2015) https://catalogue.ceh.ac.uk/documents/ce391488-1b3c-4f82-9289-4beb8b8aa7da (last access: 23 June 2017) | 15 arcsec × 15 arcsec | – | (4a–f) Average, minimum, maximum, first quartile, second quartile, and third quartile values of catchment topographic index |
| Drainage density | GRIN – Global River Network (Schneider et al., 2017) https://www.metis.upmc.fr/fr/node/375 (last access: 23 June 2017) | 7.5 arcmin × 7.5 arcmin | – | (5a–f) Average, minimum, maximum, first quartile, second quartile, and third quartile values of catchment drainage density ($km^{-1}$) |
| Dams | Global Reservoir and Dam (GRanD), version 1 (Lehner et al., 2011) http://sedac.ciesin.columbia.edu/data/set/grand-v1-dams-rev01 (last access: 23 June 2017) | 6862 datapoints storage capacity of more than 0.1 km3 | – | (6) Number of dams upstream (7) Total upstream storage volume |
| Population | Gridded Population of the World (GPW) version 4 (CIESIN, 2016) http://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-count (last access: 23 June 2017) | 30 arcsec × 30 arcsec | 2005–2014 | (8a–f) Average, minimum, maximum, first quartile, second quartile, and third quartile values of catchment population (2010) (9) 2010 Population count |
| Urbanization | Night Light Development Index (NLDI) dataset (Elvidge et al., 2012) http://www.soc-geogr.net/7/23/2012/sg-7-23-2012.html (last access: 23 June 2017) | 0.25 arcdeg × 0.25 arcdeg | 2006 | (10a–f) Average, minimum, maximum, first quartile, second quartile, and third quartile values of NLDI over catchment |
| Irrigation | Historical Irrigation Dataset (Siebert et al., 2015) https://mygeohub.org/publications/8/2 (last access: 23 June 2017) | 5 arcmin × 5 arcmin | 2005 | (11a–f) Average, minimum, maximum, first quartile, second quartile, and third quartile values of catchment Irrigated area (2005) |
| Climate type | World map of Koppen–Weiger climate classification system (Rubel and Kottek, 2010) http://koeppen-geiger.vu-wien.ac.at (last access: 23 June 2017) | 5 arcmin × 5 arcmin | 1951–2000 | (12) Type of catchment climate (Koppen–Weiger) if one type present over more than 50 % catchment area, or "No dominant type" |
| Land cover | The Climate Change Initiative Land Cover (CCI-LC) dataset http://maps.elie.ucl.ac.be/CCI/viewer/download.php (last access: 23 June 2017) | 7.5 arcsec × 7.5 arcsec | 2015 | (13) Type of catchment land cover (UN Land Cover Classification System) for 2015 if one type present over more than 50 % catchment area, or "No dominant type" |
| Lithological | The Global Lithological Map v1.0 (GLiM) dataset (Hartmann and Moosdorf, 2012) https://www.clisap.de/research/b:-climate-manifestations-and-impacts/crg-chemistry-of-natural-aqueous-solutions/global-lithological-map/ (last access: 23 June 2017) | 0.5 arcdeg × 0.5 arcdeg | – | (14) Type of catchment lithology if one type present over more than 50 % catchment area or "No dominant type" |
| Soil profile | Soil grid 250 m (Hengl et al., 2017) https://soilgrids.org (last access: 23 June 2017) | 7.5 arcsec × 7.5 arcsec | – | (15) Type of catchment soil class (World Reference Base) if one type present over more than 50 % catchment area or multiple types "No dominant type". (16a–f) Average, minimum, maximum, first quartile, second quartile, and third quartile values of weight percentage of sand over the catchment (17a–f) Average, minimum, maximum, first quartile, second quartile, and third quartile values of weight percentage of silt over the catchment (18a–f) Average, minimum, maximum, first quartile, second quartile, and third quartile values of weight percentage of clay over the catchment (19a–f) Average, minimum, maximum, first quartile, second quartile, and third quartile values of bulk content of soil over the catchment ($kg\,m^{-3}$) |

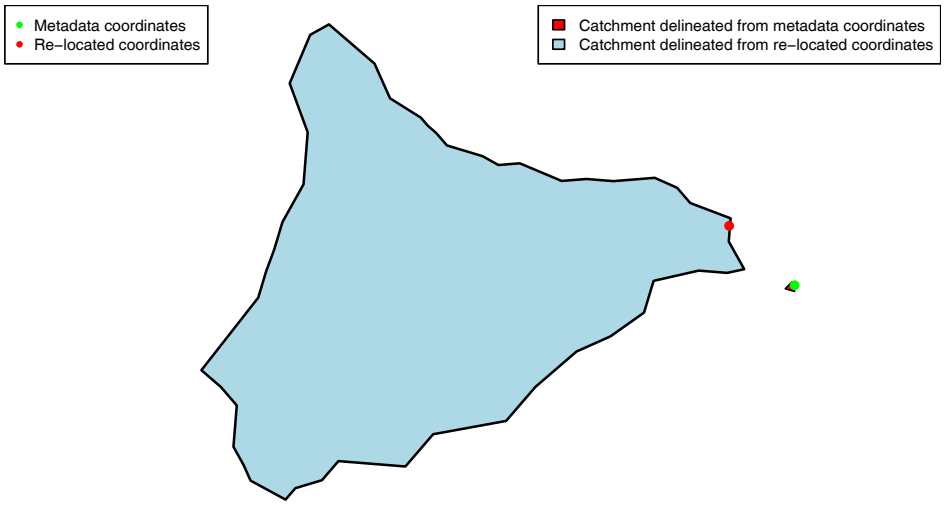


**Figure 4.** Example of improvement in quality of a catchment boundary using re-located geographical coordinates (for station AR_0000007).

density in regions with sparse streamflow observation systems (Russia, China, and Thailand, respectively). In addition, 2735 EWA stations (frozen since October 2014) were also included into GSIM as these time series have not been completely mirrored into GRDB database at the time GSIM was initiated. As these "static" time series have been frozen and no further update were provided, GSIM users are advised to use them with caution as the data may contain errors and/or have been replaced or updated.

As shown in Table 6, it is apparent that spatial coverage of the stations in the GSIM database varies significantly across continents, with North America and Europe having the greatest number of stations. Including the national databases such as MLIT (Japan), ANA (Brazil), BOM (Australia), and IWRIS (India) has significantly improved the observational network over the regions of Asia, South America, and Oceania (top panel of Fig. 5), some of which have recorded streamflow since the mid-20th century and were still operating at the time the GSIM database was initiated. This suggests that the national databases that are currently available should be given more attention in order to improve the quality and quantity of international archives.

Regarding temporal coverage, streamflow records across the globe are generally available for the second half of the 20th century (as shown in the bottom panel of Fig. 5). Regardless of missing data criteria, the number of available data gradually rises to its peak in the late 1970s to early 1980s, followed by a mild decrease in the late 1980s as also discussed by Hannah et al. (2011) and a secondary peak in the late 2000s. While the overall database has over 30 000 gauges, it is clear from Fig. 5 that from the 1960s onwards there are approximately from 10 000 to 15 000 gauges simultaneously active. This represents a significant increase in availability compared to the GRDB dataset, which had a total of approximately 9000 gauges and with a similar drop-off in available gauges depending on the filtering criteria applied.

## 5.2 Data products of GSIM

### 5.2.1 GSIM catalogue

The GSIM catalogue is designed for users to easily filter stations according to their purpose of application, and where necessary to transparently identify steps taken in the development of GSIM. The total number of 27 fields included in this document can be divided into three groups, namely the following.

1. Basic metadata. This group provides station identification, including a unique GSIM number, the name of the river, the name of the station, the elevation of the gauge, the provided geographical coordinates, and the catchment area.

2. Database merging metadata. This group of fields provides the identity of the numbers of original source(s), and if applicable the similarity metrics between duplicates.

3. Data availability metadata. This group of fields provides an overview of the data availability of each time series. These statistics were generated from the time-series data and can be used to filter station information, such as temporal coverage, data length, and the fraction of missing data.

As illustrated in Table 7, source datasets had significant gaps in the metadata, especially in cases of gauge elevation (not available in CHDP, GAME, HYDAT, BOM, and MLIT) and catchment area (not available in BOM and MLIT). In addition, the geographical coordinates of all stations were not

**Table 6.** Summary statistics of GSIM time series.

| Continent | Number of stations | Average temporal coverage (years) | Shortest record (years) | Longest record (years) | Year of earliest entry | Year of latest entry |
|---|---|---|---|---|---|---|
| Africa | 949 | 33.8 | 1 | 110 | 1903 | 2015 |
| Europe | 5778 | 40.3 | 1 | 208 | 1806 | 2016 |
| Asia | 1915 | 22.2 | 1 | 79 | 1921 | 2015 |
| North America | 15 884 | 42.9 | 1 | 156 | 1860 | 2016 |
| South America | 3449 | 29.3 | 1 | 116 | 1901 | 2016 |
| Australia and Oceania | 2984 | 31.4 | 1 | 131 | 1886 | 2016 |
| Global | 30 959 | 38.2 | 1 | 208 | 1806 | 2016 |

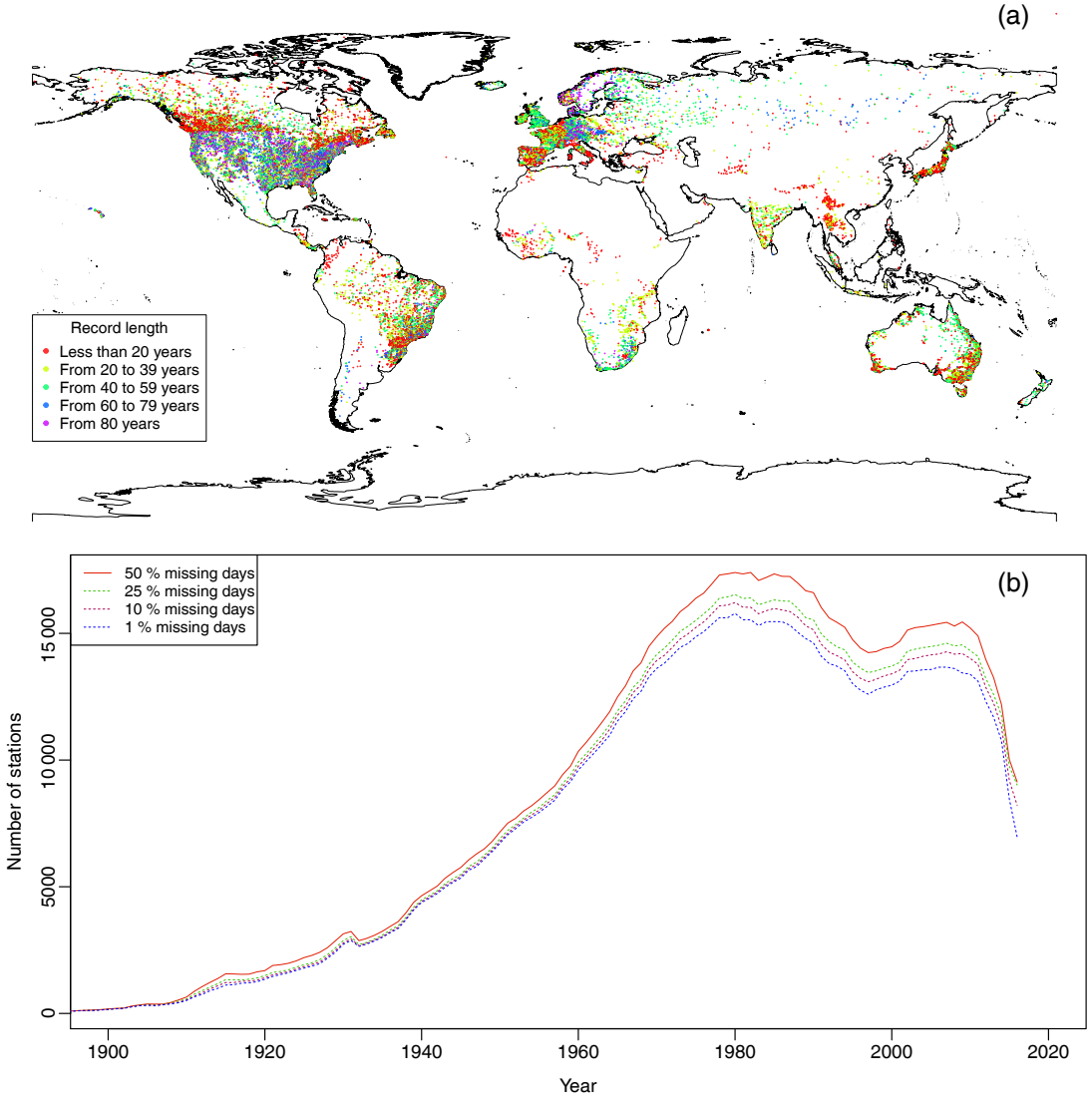


**Figure 5.** Availability of GSIM time series. **(a)** illustrates the length of record at each station, and **(b)** illustrates the number of available time series over time for four different missing data criteria.

**Table 7.** The percentage of stations accompanied by all basic metadata.

| Dataset | Station ID | River name | Station name | Latitude | Longitude | Altitude | Catchment area |
|---|---|---|---|---|---|---|---|
| ADF | 100 | 100 | 100 | 100 | 100 | 96.2 | 99.3 |
| ANA | 100 | 99.9 | 100 | 100 | 100 | 69 | 99 |
| ARCTICNET | 100 | 100 | 100 | 99.3 | 99.3 | 99.3 | 100 |
| BOM | 100 | 100 | 100 | 100 | 100 | 0 | 0 |
| CHDP | 100 | 99.4 | 100 | 100 | 100 | 0 | 84 |
| EWA | 100 | 100 | 100 | 100 | 100 | 98.5 | 94.5 |
| GAME | 100 | 100 | 100 | 100 | 100 | 0 | 100 |
| GRDB | 100 | 100 | 100 | 100 | 100 | 67 | 100 |
| HYDAT | 100 | 100 | 100 | 100 | 100 | 0 | 85.8 |
| MLIT | 100 | 100 | 100 | 100 | 100 | 0 | 0 |
| USGS | 100 | 100 | 100 | 100 | 100 | 93.7 | 25.5 |
| WRIS | 100 | 100 | 100 | 100 | 100 | 81.6 | 97.4 |
| GSIM | 100 | 99.9 | 100 | 99.9 | 99.9 | 50.4 | 73.8 |

**Table 8.** Percentages of available catchment-scale characteristics.

| Catchment characteristics | Number of stations | Availability percentage |
|---|---|---|
| Climate classification | 30 773 | 99.5 |
| Drainage density | 29 574 | 95.6 |
| Elevation | 30 932 | 99.9 |
| Irrigation area | 30 857 | 99.7 |
| Land cover classification | 30 888 | 99.8 |
| Lithology type | 30 154 | 97.5 |
| Nightlight Development Index | 23 096 | 74.7 |
| Population count | 30 894 | 99.9 |
| Population density | 30 800 | 99.6 |
| Slope | 30 862 | 99.8 |
| Soil bulk density | 30 812 | 99.6 |
| Soil classification | 30 764 | 99.4 |
| Clay content | 30 768 | 99.5 |
| Clay content | 30 695 | 99.2 |
| Silt content | 30 828 | 99.7 |
| Topographic index | 30 725 | 99.3 |

correctly recorded for all stations, with 24 removed as having suspect locations and 4871 shifted coordinates as part of the procedure for aligning catchment outlets with reported catchment areas.

### 5.2.2 Quality of catchment boundary

The catchment boundary is the second metadata product that is available through GSIM. Of all GSIM stations, 12 150 (39 %) were not associated with any information about drainage areas (including all MLIT and BOM stations); thus, a "Caution" flag is attached to upstream catchments of these stations. Another 24 stations with suspected geographical coordinates of stations were removed, and the final 18 785 stations were processed to identify the "best outlet" location to represent the outlet for delineating upstream catchments. The distribution and quality of the delineated catchments of these stations are provided in Fig. 6 (figures at continental scale are also provided as a Supplement).

As illustrated in the top panel, "Caution" catchments using "best" outlets (identified using the method outlined in Sect. 4.2) are generally located across all GSIM regions. However, the "Caution" flag appears more frequently over regions above 60° N. Further checks would be required to improve the association of catchment boundaries with stations. Unfortunately, the biggest caveat that applies to the GSIM database, as with any global database, is that the metadata were collated from a number of sources with varying standards of documentation and quality assurance and with limited capacity for additional checking other than automated procedures. Therefore, there is likely to be a non-trivial degree of error in the metadata for both geographical location and drainage area. Another issue that may lead to unreliable results of the delineation process is error in the DEM products. This potential error has been documented (Lehner, 2012; Lehner et al., 2006), and lower-quality DEM products generally exist for regions above 60° N due to the lower quality of the original elevation products used to derive the DEM datasets. Another note for the use of delineated catchments is that very small catchments (area less than 50 $km^2$) should be handled with care, as the "best" outlets could be located incorrectly while still delivering "acceptable" discrepancies as part of the automated procedure.

Nonetheless, the quality of delineated catchments is quite positive (as illustrated in the lower panels of Fig. 6). Of all 18 785 catchments that had reported drainage area in the GSIM catalogue, 68.25, 11.8, and 15.92 % of catchments have "High" quality (area discrepancy of less than 5 %), "Medium" quality (area discrepancy from 5 % to less than 10 %), and "Low" quality (area discrepancy from 10 to less than 50 %), respectively, while there are only 4.03 % catch-

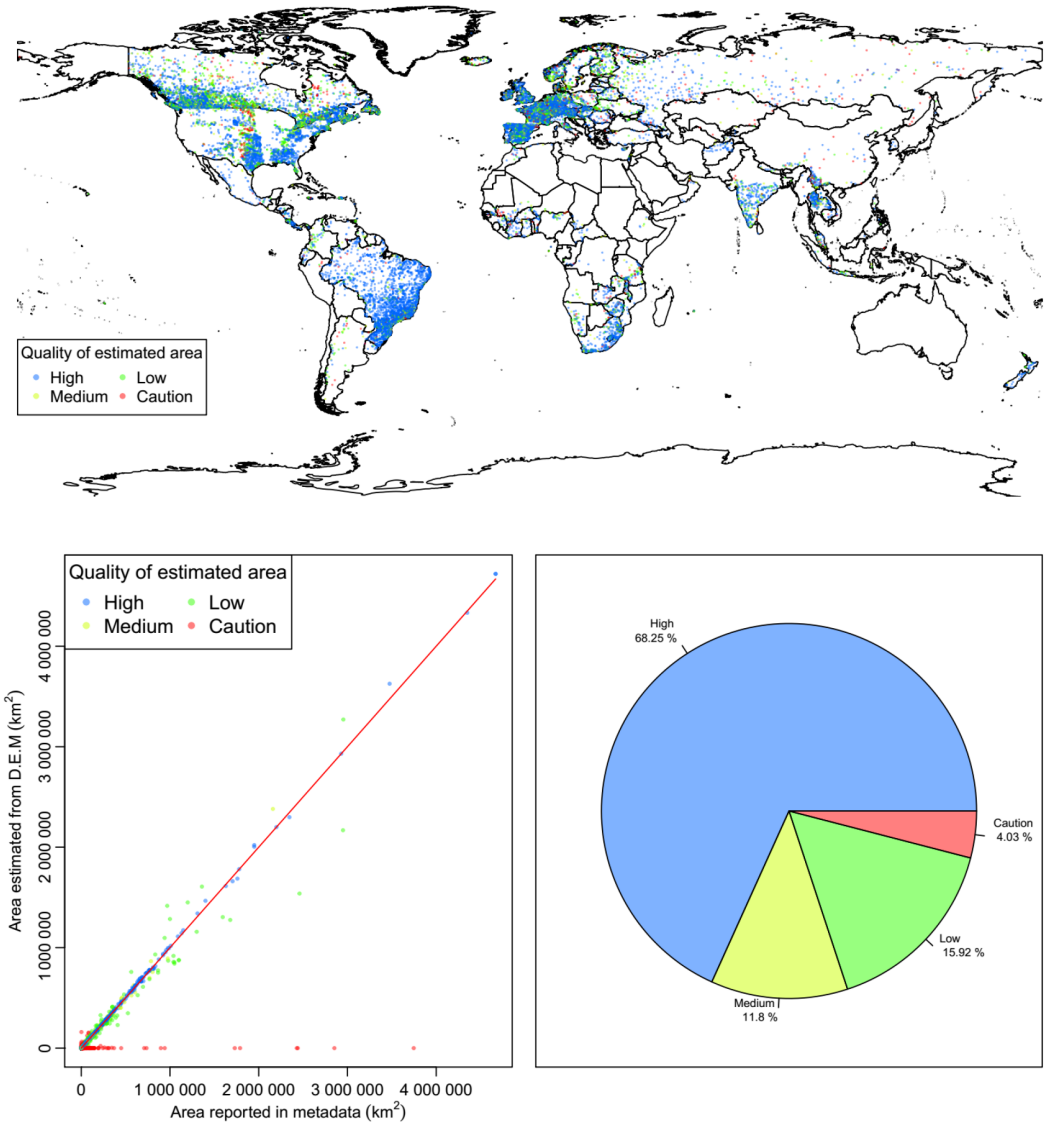


**Figure 6.** Quality of the delineated catchment boundary according to the categories of high, medium, low, and caution identified in Sect. 4.2 (for 18 785 stations that have reported drainage area and reasonable geographical coordinates).

ments with "Caution" quality (area discrepancy of more than or equal to 50 %).

### 5.2.3 Catchment-scale characteristics

The final data product that has been made available is the auxiliary information extracted from 12 global coverage datasets representing many characteristics associated with GSIM stations. Overall, the spatial coverage of original data products (mostly satellite-based is quite good (see Table 8), with just a small fraction of catchments (less than 10 %) that have more than 30 % of their areas not covered by these datasets. The exception is the Nightlight Development Index (NLDI – computed from the 2006 Nightlights dataset, Ziskin et al., 2010, and the 2006 Landscan gridded population, Bhaduri et al., 2002). This dataset does not have approximately 25.3 % of catchments covered, for more than 70 % of their areas.

It is important to note that while these catchment-scale characteristics are consistent products available for all stations, documentation for the original source data should be consulted during application to appreciate the limitations and appropriateness of each variable. For example, the GRanD database is not exhaustive of all dams worldwide and there can be ambiguities over the affiliated dates (e.g. whether they represent conception, construction, or commissioning). Furthermore, the extent of the overlapping period between temporal coverage of streamflow time series and remote sensing based datasets needs to be carefully assessed in cause–

effect studies. Similarly, it is likely that there will be updated or new data gridded datasets available over time so that applications should consider the appropriateness of the information used. The availability of metadata products emerging from the GSIM project demonstrates the possibility of using reported global data products to extract catchment-scale characteristics associated with each station with reasonable quality, enabling many potential applications from this rich information.

## 6 Data availability

The data described in this paper are available as a compressed zip archive containing (i) a readme file, (ii) metadata of all GSIM stations obtained from original data sources and time series, (iii) quality of catchment boundary and catchment characteristics extracted from 12 global data products, (iv) a list of stations with suspect geographical coordinates, and (v) catchment boundaries for 30 935 stations that have a reasonable geographical location.

The data can be freely downloaded at PANGEA data depository https://doi.pangaea.de/10.1594/PANGAEA.887477 (Do et al., 2018). The uploaded zip archive contains two directories and one README.txt file. The readme file provides a detailed description of the data. The "GSIM_catalogue" directory contains the metadata of all GSIM stations and a list of stations with suspect geographical coordinates. The "GSIM_catchments" directory contains shapefiles for 30 935 stations.

## 7 Conclusions

In situ observations of daily streamflow with global coverage are crucial to understanding large-scale freshwater resources that are fundamental for societal development. The GSIM archive, designed as an expansion of the GRDB database, has demonstrated the possibility of significantly improving the coverage and density of the global streamflow observational datasets using free-to-access databases. The development of the GSIM database was not possible without the tremendous investment in the production and ongoing maintenance of original data sources of GSIM. This fact emphasizes the key role of data authorities and international initiatives in enabling advances in large-scale hydrology by making data publicly available to the community.

While the activities of GSIM have been extensive in searching out and collating databases, they are by no means exhaustive (e.g. since submission we have been notified of additional potential candidates for inclusion such as the Mekong River Commission database, Chile national water database, and Argentina national water database). It is the authors' intention that this project will stimulate further efforts toward the development of coordinated and consistent representation of global streamflow observations. For this reason, the process of developing the archive was designed with automation in mind. With the exception of needing to visually inspect some cases of duplicated time series, the archive was automated using scripts in the R and Python programming languages.

Although the GSIM database was compiled from data sources that can be obtained free of charge via a data portal or by submitting written requests to data authorities, there are some strict conditions related to the redistribution of unprocessed data. Therefore, it is impossible to make the whole GSIM collection publicly available. In addition, with the main aim of harvesting as much data as possible, the GSIM database is not focused on collecting high-quality datasets such as referenced hydrological networks that are available in many countries (Whitfield et al., 2012), and thus the data quality may vary significantly across the available time series. To address these limitations and increase the usefulness of the GSIM database, we conducted a set of quality checking procedures for all GSIM time series. These quality-assured records were then used to produce a dedicated set of indices capturing important aspects of the daily dynamics from GSIM time series, and to explore potential applications of GSIM in large-scale hydrology. Detailed information about this work and associated distributed data is described in the second part of our series on GSIM (Gudmundsson et al., 2018a, b).

With the GSIM archive and production information made publicly available in a transparent manner, this project serves the broader hydrology community with improved coverage and quality of streamflow information. This project has yielded a significant increase in the availability of streamflow observations through the process of collating readily accessed online data, and with ongoing efforts there will be opportunities for further extension. Streamflow observations represent an underutilized resource, in part due to access limitations, but also due to challenges in accounting for human impacts in the observed record. These challenges notwithstanding, ongoing advances in global-scale hydrological models and ever-increasing access to remote-sensed products indicate that wider access to streamflow data has the potential to significantly enhance our knowledge of global water resources.

**The Supplement related to this article is available online at https://doi.org/10.5194/essd-10-765-2018-supplement.**

**Competing interests.** The authors declare that they have no conflict of interest.

**Acknowledgements.** The authors would like to express their appreciation to all the national agencies and institutions that made the streamflow data available for this study. We would like to thank

Sonia I. Seneviratne for her discussions and support on the collation of the GSIM archive. Hong Xuan Do receives financial support from the Australia Award Scholarship (AAS). Seth Westra's time was supported by Australian Research Council Discovery project DP150100411. The authors also wish to thank two reviewers for their constructive comments and suggestions. The authors would like to express their sincere thanks to Danlu Guo for her support in collecting the MLIT database. This work was supported with supercomputing resources provided by the Phoenix HPC service at the University of Adelaide.

Edited by: David Carlson
Reviewed by: Wolfgang Grabs and one anonymous referee

## References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci., 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017.

Arsenault, R., Bazile, R., Ouellet Dallaire, C., and Brissette, F.: CANOPEX: A Canadian hydrometeorological watershed database, Hydrol. Process., 30, 2734–2736, 2016.

Barr, I. D. and Clark, C. D.: An updated moraine map of Far NE Russia, J. Maps, 8, 431–436, 2012.

Bhaduri, B., Bright, E., Coleman, P., and Dobson, J.: LandScan, Geoinformatics, 5, 34–37, 2002.

Bierkens, M. F. P.: Global hydrology 2015: State, trends, and directions, Water Resour. Res., 51, 4923–4947, 2015.

Bierkens, M. F. P., Bell, V. A., Burek, P., Chaney, N., Condon, L. E., David, C. H., de Roo, A., Döll, P., Drost, N., Famiglietti, J. S., Flörke, M., Gochis, D. J., Houser, P., Hut, R., Keune, J., Kollet, S., Maxwell, R. M., Reager, J. T., Samaniego, L., Sudicky, E., Sutanudjaja, E. H., van de Giesen, N., Winsemius, H., and Wood, E. F.: Hyper-resolution global hydrological modelling: what is next?, Hydrol. Process., 29, 310–320, 2015.

Burn, D. H., Hannaford, J., Hodgkins, G. A., Whitfield, P. H., Thorne, R., and Marsh, T.: Reference hydrologic networks II. Using reference hydrologic networks to assess climate-driven changes in streamflow, Hydrolog. Sci. J., 57, 1580–1593, 2012.

Christen, P.: Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection, Springer Science & Business Media, 2012.

CIESIN: Gridded Population of the World, Version 4 (GPWv4): Population Density Adjusted to Match 2015 Revision UN WPP Country Totals, NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, NY, 2016.

Do, H. X., Westra, S., and Michael, L.: A global-scale investigation of trends in annual maximum streamflow, J. Hydrol., 552, 28–43, https://doi.org/10.1016/j.jhydrol.2017.06.015, 2017.

Do, H. X., Gudmundsson, L., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive – Part 1: Station catalog and Catchment boundary, PANGAEA, https://doi.org/10.1594/PANGAEA.887477, 2018.

Elvidge, C. D., Baugh, K. E., Anderson, S. J., Sutton, P. C., and Ghosh, T.: The Night Light Development Index (NLDI): a spatially explicit measure of human development from satellite data, Soc. Geogr., 7, 23–35, 2012.

Falcone, J. A.: GAGES-II: Geospatial attributes of gages for evaluating streamflow, US Geological Survey, 2011.

Fekete, B. M., Vörösmarty, C., and Grabs, W.: Global Composite Runoff Fields on Observed River Discharge and Simulated Water Balances/Water System Analysis Group. University of New Hampshire, and Global Runoff Data Centre, Koblenz, Federal Institute of Hydrology (BfG), Koblenz, Germany, Federal Institute of Hydrology (BfG), 2002a.

Fekete, B. M., Vörösmarty, C. J., and Grabs, W.: High-resolution fields of global runoff combining observed river discharge and simulated water balances, Global Biogeochem. Cy., 16, 15-11–15-10, 2002b.

Fekete, B. M., Robarts, R. D., Kumagai, M., Nachtnebel, H.-P., Odada, E., and Zhulidov, A. V.: Time for in situ renaissance, Science, 349, 685–686, 2015.

Fredin, O., Rubensdotter, L., van Welden, A., Larsen, E., and Lysså, A.: Distribution of ice marginal moraines in NW Russia, J. Maps, 8, 236–241, 2012.

GRDC: Report of the Twelfth Meeting of the GRDC Steering Committee, Koblenz, Germany, 18–19 June 2014, Global Runoff Data Centre (GRDC), Koblenz, Germany, 23 pp., 2015.

Gudmundsson, L. and Seneviratne, S. I.: Towards observation-based gridded runoff estimates for Europe, Hydrol. Earth Syst. Sci., 19, 2859–2879, https://doi.org/10.5194/hess-19-2859-2015, 2015.

Gudmundsson, L. and Seneviratne, S. I.: Observation-based gridded runoff estimates for Europe (E-RUN version 1.1), Earth Syst. Sci. Data, 8, 279–295, https://doi.org/10.5194/essd-8-279-2016, 2016.

Gudmundsson, L., Tallaksen, L. M., Stahl, K., Clark, D. B., Dumont, E., Hagemann, S., Bertrand, N., Gerten, D., Heinke, J., Hanasaki, N., Voss, F., and Koirala, S.: Comparing Large-Scale Hydrological Model Simulations to Observed Runoff Percentiles in Europe, J. Hydrometeorol., 13, 604–620, 2012.

Gudmundsson, L., Seneviratne, S. I., and Zhang, X.: Anthropogenic climate change detected in European renewable freshwater resources, Nat. Clim. Change, 7, 813–816, https://doi.org/10.1038/nclimate3416, 2017.

Gudmundsson, L., Do, H. X., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive (GSIM) – Part 2: Quality control, time-series indices and homogeneity assessment, Earth Syst. Sci. Data, 10, 787–804, https://doi.org/10.5194/essd-10-787-2018, 2018a.

Gudmundsson, L., Do, H. X., Leonard, M., and Westra, S.: The Global Streamflow Indices and Metadata Archive (GSIM) – Part 2: Time Series Indices and Homogeneity Assessment, PANGAEA, https://doi.org/10.1594/PANGAEA.887470, 2018b.

Hannaford, J. and Marsh, T.: An assessment of trends in UK runoff and low flows using a network of undisturbed catchments, Int. J. Climatol., 26, 1237–1253, 2006.

Hannah, D. M., Demuth, S., van Lanen, H. A. J., Looser, U., Prudhomme, C., Rees, G., Stahl, K., and Tallaksen, L. M.: Large-scale river flow archives: importance, current status and future needs, Hydrol. Process., 25, 1191–1200, 2011.

Hartmann, J. and Moosdorf, N.: The new global lithological map database GLiM: A representation of rock properties at the Earth surface, Geochem. Geophy. Geosy., 13, Q12004, https://doi.org/10.1029/2012GC004370, 2012.

Henck, A. C., Montgomery, D. R., Huntington, K. W., and Liang, C.: Monsoon control of effective discharge, Yunnan and Tibet, Geology, 38, 975–978, 2010.

Hengl, T., de Jesus, J. M., Heuvelink, G. B., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., and Bauer-Marschallinger, B.: SoilGrids250m: Global gridded soil information based on machine learning, PloS one, 12, e0169748, https://doi.org/10.1371/journal.pone.0169748, 2017.

Hodgkins, G. A., Whitfield, P. H., Burn, D. H., Hannaford, J., Renard, B., Stahl, K., Fleig, A. K., Madsen, H., Mediero, L., Korhonen, J., Murphy, C., and Wilson, D.: Climate-driven variability in the occurrence of major floods across North America and Europe, J. Hydrol., 552, 704–717, 2017.

Huang, S., Kumar, R., Flörke, M., Yang, T., Hundecha, Y., Kraft, P., Gao, C., Gelfan, A., Liersch, S., Lobanova, A., Strauch, M., van Ogtrop, F., Reinhardt, J., Haberlandt, U., and Krysanova, V.: Evaluation of an ensemble of regional hydrological models in 12 large-scale river basins worldwide, Climatic Change, 141, 381–397, https://doi.org/10.1007/s10584-016-1841-8, 2016.

Jenson, S. K. and Domingue, J. O.: Extracting topographic structure from digital elevation data for geographic information-system analysis, Photogramm. Eng. Rem. S., 54, 1593–1600, 1988.

Kundzewicz, Z. W., Plate, E. J., Rodda, H. J., Rodda, J. C., Schellnhuber, H. J., and Strupczewski, W. G.: Changes in flood risk–setting the stage. In: Changes in flood risk in Europe, CRC Press, 2012.

Kundzewicz, Z. W., Kanae, S., Seneviratne, S. I., Handmer, J., Nicholls, N., Peduzzi, P., Mechler, R., Bouwer, L. M., Arnell, N., Mach, K., Muir-Wood, R., Brakenridge, G. R., Kron, W., Benito, G., Honda, Y., Takahashi, K., and Sherstyukov, B.: Flood risk and climate change: global and regional perspectives, Hydrolog. Sci. J., 59, 1–28, 2013.

Lehner, B.: Derivation of watershed boundaries for GRDC gauging stations based on the HydroSHEDS drainage network, Report 41 in the GRDC Report Series, available at: http://www.bafg.de/GRDC/EN/02_srvcs/24_rprtsrs/report_41.html?nn=201764 (last access: 23 June 2017), 2012.

Lehner, B. and Grill, G.: Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems, Hydrol. Process., 27, 2171–2186, 2013.

Lehner, B., Liermann, C. R., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., Frenken, K., and Magome, J.: Global reservoir and dam (grand) database, Technical Documentation, Version, 1, 2011.

Lehner, B., Verdin, K., and Jarvis, A.: HydroSHEDS technical documentation, version 1.0, World Wildlife Fund US, Washington, DC, 1–27, 2006.

Lehner, B., Verdin, K., and Jarvis, A.: New Global Hydrography Derived From Spaceborne Elevation Data, Eos, Transactions American Geophysical Union, 89, 93–94, 2008.

Lindsay, J. B., Rothwell, J. J., and Davies, H.: Mapping outlet points used for watershed delineation onto DEM-derived stream networks, Water Resour. Res., 44, W08442, https://doi.org/10.1029/2007WR006507, 2008.

Lucas, C. and Tingley, D.: translateR, available at: https://cran.r-project.org/web/packages/translateR/index.html (last access: 23 June 2017), 2016.

Marthews, T. R., Dadson, S. J., Lehner, B., Abele, S., and Gedney, N.: High-resolution global topographic index values for use in large-scale hydrological modelling, Hydrol. Earth Syst. Sci., 19, 91–104, https://doi.org/10.5194/hess-19-91-2015, 2015.

Merz, B., Kundzewicz, Z., Delgado, J., Hundecha, Y., and Kreibich, H.: Detection and attribution of changes in flood hazard and risk, Changes in flood risk in Europe, IAHS Special Publication, 10, 435–458, 2012.

Milly, P. C. D., Wetherald, R. T., Dunne, K., and Delworth, T. L.: Increasing risk of great floods in a changing climate, Nature, 415, 514–517, 2002.

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., Stouffer, R. J., Dettinger, M. D., and Krysanova, V.: On Critiques of "Stationarity is Dead: Whither Water Management?", Water Resour. Res., 51, 7785–7789, 2015.

Nelson, B.: Data sharing: Empty archives, Nature, 461, 160–163, https://doi.org/10.1038/461160a, 2009.

Rubel, F. and Kottek, M.: Observed and projected climate shifts 1901–2100 depicted by world maps of the Köppen-Geiger climate classification, Meteorol. Z., 19, 135–141, 2010.

Schaake, J., Cong, S., and Duan, Q.: The US MOPEX data set, IAHS publication, 307, 9–28, 2006.

Schneider, A., Jost, A., Coulon, C., Silvestre, M., Théry, S., and Ducharne, A.: Global-scale river network extraction based on high-resolution topography and constrained by lithology, climate, slope, and observed drainage density, Geophys. Res. Lett., 44, 2773–2781, 2017.

Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C. M., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., and Rahimi, M.: Changes in climate extremes and their impacts on the natural physical environment, Managing the risks of extreme events and disasters to advance climate change adaptation, 109–230, 2012.

Siebert, S., Kummu, M., Porkka, M., Döll, P., Ramankutty, N., and Scanlon, B. R.: A global data set of the extent of irrigated land from 1900 to 2005, Hydrol. Earth Syst. Sci., 19, 1521–1545, https://doi.org/10.5194/hess-19-1521-2015, 2015.

Sil, A. and Sitharam, T.: Detection of Local Site Conditions in Tripura and Mizoram Using the Topographic Gradient Extracted from Remote Sensing Data and GIS Techniques, Nat. Hazards Rev., 18, 04016009, https://doi.org/10.1061/(ASCE)NH.1527-6996.0000228, 2016.

Vörösmarty, C. J., Moore, B., Grace, A. L., Gildea, M. P., Melillo, J. M., Peterson, B. J., Rastetter, E. B., and Steudler, P. A.: Continental scale models of water balance and fluvial transport: an application to South America, Global Biogeochem. Cy., 3, 241–265, 1989.

Wanders, N. and Wada, Y.: Decadal predictability of river discharge with climate oscillations over the 20th and early 21st century, Geophys. Res. Lett., 42, 10689–610695, 2015.

Ward, P. J., Jongman, B., Weiland, F. S., Bouwman, A., van Beek, R., Bierkens, M. F., Ligtvoet, W., and Winsemius, H. C.: Assessing flood risk at the global scale: model setup, results, and sensitivity, Environ. Res. Lett., 8, 044019, https://doi.org/P10.1088/1748-9326/8/4/044019, 2013.

Ward, P. J., Eisner, S., Flörke, M., Dettinger, M. D., and Kummu, M.: Annual flood sensitivities to El Niño-Southern Oscilla-

tion at the global scale, Hydrol. Earth Syst. Sci., 18, 47–66, https://doi.org/10.5194/hess-18-47-2014, 2014.

Ward, P. J., Jongman, B., Salamon, P., Simpson, A., Bates, P., De Groeve, T., Muis, S., de Perez, E. C., Rudari, R., Trigg, M. A., and Winsemius, H. C.: Usefulness and limitations of global flood risk models, Nat. Clim. Change, 5, 712–715, 2015.

Whitfield, P. H., Burn, D. H., Hannaford, J., Higgins, H., Hodgkins, G. A., Marsh, T., and Looser, U.: Reference hydrologic networks I. The status and potential future directions of national reference hydrologic networks for detecting trends, Hydrolog. Sci. J., 57, 1562–1579, 2012.

Wood, E. F., Roundy, J. K., Troy, T. J., van Beek, L. P. H., Bierkens, M. F. P., Blyth, E., de Roo, A., Döll, P., Ek, M., Famiglietti, J., Gochis, D., van de Giesen, N., Houser, P., Jaffé, P. R., Kollet, S., Lehner, B., Lettenmaier, D. P., Peters-Lidard, C., Sivapalan, M., Sheffield, J., Wade, A., and Whitehead, P.: Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water, Water Resour. Res., 47, W05301, https://doi.org/10.1029/2010WR010090, 2011.

Yamazaki, D., Trigg, M. A., and Ikeshima, D.: Development of a global ∼ 90 m water body map using multi-temporal Landsat images, Remote Sens. Environ., 171, 337–351, 2015.

Ziskin, D., Baugh, K., Hsu, F.-C., and Elvidge, C. D.: Methods used for the 2006 radiance lights, Proceedings of the Asia-Pacific Advanced Network, 30, 131–142, 2010.