

A Synthesis of Global Streamflow characteristics, Hydrometeorology, and catchment Attributes (GSHA) for Large Sample River-Centric Studies

Ziyun Yin¹, Peirong Lin^{1,2*}, Ryan Riggs³, George H. Allen⁴, Xiangyong Lei¹, Ziyang Zheng^{5,6}, Siyu Cai⁷

1. Institute of Remote Sensing and GIS, School of Earth and Space Sciences, Peking University
2. International Research Centre for Big Data for Sustainable Development Goals, Beijing, China
3. Department of Geography, Texas A&M University, Texas, USA
4. Department of Geosciences, Virginia Polytechnic Institute and State University, Virginia, USA
5. Key Laboratory of Regional Climate-Environment Research for Temperate East Asia, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China
6. University of Chinese Academy of Sciences, Beijing, China
7. State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research, Beijing, China

* *Correspondence to:* Peirong Lin (peironglinlin@pku.edu.cn)

Revised manuscript submitted to *ESSD*, November 15th, 2023

Abstract

Our understanding and predictive capability of streamflow processes largely rely on high-quality datasets that depict a river's upstream basin characteristics. Recent proliferation of large sample hydrology (LSH) datasets has promoted model parameter estimation and data-driven analyses of the hydrological processes worldwide, yet existing LSH is still insufficient in terms of sample coverage, uncertainty estimates, and dynamic descriptions of anthropogenic activities. To bridge the gap, we contribute the Synthesis of Global Streamflow characteristics, Hydrometeorology, and catchment Attributes (GSHA) to complement existing LSH datasets, which covers 21,568 watersheds from 13 agencies for as long as 43 years based on discharge observations scraped from web. In addition to annual and monthly streamflow indices, each basin's daily meteorological variables (i.e., precipitation, 2 m air temperature, longwave/shortwave radiation, wind speed, actual and potential evapotranspiration), daily-weekly water storage terms (i.e., snow water equivalence, soil moisture, groundwater percentage), and yearly dynamic descriptors of the land surface characteristics (i.e., urban/cropland/forest fractions, leaf area index, reservoir storage and degree of regulation) are also provided by combining openly available remote sensing and reanalysis datasets. The uncertainties of all meteorological variables are estimated with independent data sources. Our analyses reveal the following insights: (i) the meteorological data uncertainties vary across variables and geographical regions, and the revealed pattern should be accounted for by LSH users, (ii) ~6% watersheds shifted between human managed and natural states during 2001-2015, e.g., basins with environmental recovery projects in Northeast China, which may be useful for hydrologic analysis that takes the changing land surface characteristics into account, and (iii) GSHA watersheds showed a more widespread declining trend in runoff coefficient than an increasing trend, pointing towards critical water availability issues. Overall, GSHA is expected to serve hydrological model parameter estimation and data-driven analyses as it continues to improve. GSHA v1.1 can be accessed at

41 <https://doi.org/10.5281/zenodo.8090704> and <https://doi.org/10.5281/zenodo.10127757>. (Yin et al.,
42 2023).

43 1 Introduction

44 Climate change has posed profound challenges to the management of freshwater resources,
45 specifically riverine floods or water shortages (AghaKouchak et al., 2020; Thackeray et al., 2022).
46 The urgent need for flood and drought forecasting, water resources planning and management, all
47 call for high-quality streamflow predictions for basins worldwide to analyse global terrestrial water
48 conditions in a systematic view (Burgess, 1998). The scarcity of hydrological observations has
49 brought challenges to these predictions (Belvederesi et al., 2022; Hrachowitz et al., 2013), thus the
50 development of computer models that allow for “modelling everything everywhere” (Beven &
51 Alcock, 2012) constitutes the backbone of hydrological studies. Existing studies have used
52 physically-based and data-driven models for streamflow simulation (Lin et al., 2018; Nandi &
53 Reddy, 2022; Zhang et al., 2020), with efforts to improve accuracy of prediction by combining both
54 (Cho & Kim, 2022; Razavi & Coulibaly, 2013). Yet the prediction of the magnitude, timing, and
55 trend of critical streamflow characteristics are still subject to multiple sources of errors and
56 uncertainties (Bourdin et al., 2012; Brunner et al., 2021).

57 Streamflow (Q) can be represented by the simple water balance equation involving
58 precipitation (P), evapotranspiration (ET), and water storage terms (S) denoted as $Q = P - ET - \Delta S$,
59 yet influencing factors of these components could bring uncertainties that cascade downstream.
60 Starting from the model assumptions to the data used to represent climate, soil water, ice cover,
61 topography and land use, as well as the less well-known processes such as human perturbations and
62 sub-surface flows (Benke et al., 2008; Wilby & Dessai, 2010), these complications impede our
63 understanding of streamflow processes across scales, which also limits the modelling and predictive
64 capability for streamflow. Thus, reducing the predictive uncertainties requires high-quality data with
65 massive samples capable of depicting each of the water balance components, as well as the natural
66 and anthropogenic factors involved (Gupta et al., 2014).

67 Efforts have been made to address the need for such kind of high-quality datasets on watershed-
68 scale hydro-climate and environmental conditions during the past couple of decades. One of the
69 earliest was the most widely used dataset generated for the Model Parameter Estimation Experiment
70 (MOPEX) project aimed at better hydrological modelling (Duan et al., 2006). Historical hydro-
71 meteorological data and land surface characteristics for over 400 hydrologic basins in the United
72 States were provided, which was fundamental to the progress in large sample hydrology (LSH)
73 (Addor et al., 2020; Schaake et al., 2006). Later the dataset was expanded to 671 catchments in the
74 contiguous United States (CONUS) and benchmarked by model results (Newman et al., 2015).
75 Based on these studies, the Catchment Attributes and Meteorology for Large-sample Studies
76 (CAMELS) dataset was developed, providing comprehensive and updated data on topography,
77 climate, streamflow, land cover, soil, and geology attributes for each catchment (Addor et al., 2017).
78 The CONUS CAMELS dataset soon became influential in LSH and has since inspired researchers
79 from Australia (Fowler et al., 2021), Europe (Coxon et al., 2020; Delaigue et al., 2022; Klingler et

80 al., 2021), South America (Alvarez-Garreton et al., 2018; Chagas et al., 2020), and China (Hao et al., 2021) to contribute their regional CAMELS. Another comprehensive regional LSH dataset for
81 North America named the Hydrometeorological Sandbox - École de Technologies Supérieure
82 (HYSETS) dataset, was also developed with larger sample size (14425 watersheds) and richer data
83 sources compared with the CAMELS (Arsenault et al., 2020).
84

85 While these datasets are reliable data sources for regional studies, attempts on building global
86 datasets have become the new norm in the era of big data to boost our analytical and modelling
87 capability for the terrestrial hydrological processes. The HydroATLAS dataset integrated indices of
88 hydrology, physiography, climate, land cover, soil, geology, and anthropogenic activity attributes
89 for 8.5 million global river reaches (Lehner et al., 2022; Linke et al., 2019). A recent work combined
90 a series of CAMELS datasets with HydroATLAS attributes into a new global community dataset on
91 the cloud named Caravan, with dynamic hydro-climate variables and comprehensive static
92 catchment attributes extracted on 6830 watersheds (Kratzert et al., 2023), which represents by far
93 the most comprehensive synthesis of existing CAMELS. Another global-scale effort, the Global
94 Streamflow Indices and Metadata archive (GSIM), incorporated dynamic streamflow indices and
95 attribute metadata for topography, climate type, land cover, etc., for over 35000 gauges (Do et al.,
96 2018; Gudmundsson et al., 2018), and the streamflow indices were updated to allow for trend
97 analysis (Chen et al., 2023). A recent study filled in the discontinuity and latency of gauge records,
98 and provided streamflow for over 45,000 gauges with improved data quality (Riggs et al., 2023).
99 These global-scale datasets have been widely used in data-driven machine learning models (Kratzert
100 et al., 2019a, 2019b; Ren et al., 2020), physical hydrological models (Aerts et al., 2022; Clark et al.,
101 2021), and parameter estimation and regionalization studies (Addor et al., 2018; Fang et al., 2022).
102

103 Although the flourishing of LSH datasets has promoted comparative hydrological studies
104 (Kovács, 1984) and large-scale hydrological modeling and analysis efforts, several challenges are
105 still standing in the way of realizing the full potential of LSH. As briefly outlined in a recent review
106 by Addor et al. (2020), current LSH datasets lack common standards, metadata and uncertainty
107 estimates, and are insufficient in characterising human interventions. More specifically, the
108 following major critical aspects still need attention from the LSH developers, which we attempt to
109 address with GSHA (Yin et al., 2023). First, the majority of current datasets (especially those at a
110 global scale) incorporated only one data source for each variable, while earth observations,
111 reanalysis, satellite-based estimates are subject to uncertainties (Merchant et al., 2017; Ukhurebor
112 et al., 2020). These uncertainties were rarely represented and may bring difficulties to the
113 regionalization of model parameters (Beck et al., 2016), while also resulting in inconsistent
114 conclusions. Second, anthropogenic activities including land use and land cover (LULC) changes,
115 dam and reservoir building, etc., are critical drivers of shifts in streamflow statistical moments
116 (Niraula et al., 2015). However, historical time series of watershed human modifications were rarely
117 included in LSH datasets, which is particularly problematic for regions with rapid economic growth.
118 Finally, although the most recent Caravan provided hydroclimate data for global watersheds, the
119 samples are limited to the existing regional CAMELS which Caravan synthesizes. Therefore, plenty
120 of room is left to increase data sample size and spatial coverage by revisiting the streamflow data
121 acquisition process in a more comprehensive way.

122 To complement existing LSH datasets, we contribute the first version of a synthesis of Global
123 Streamflow characteristics, Hydrometeorology, and catchment Attributes (GSHA v_1.0) for large-
sample river-centric studies. GSHA features the following characteristics:

- 124 ● Updated physical and anthropogenic descriptors of global rivers, covering streamflow
125 characteristics, hydrometeorological variables, and land use land cover changes for 21568
126 watersheds derived from gauged streamflow records from 13 agencies.
- 127 ● Streamflow indices for data scarce regions, including those derived from 263 gauges in
128 China, are included.
- 129 ● Extended temporal coverage for as long as 43 years (1979-2021), which varies regionally.
- 130 ● Uncertainty estimates for the meteorological variables.
- 131 ● Dynamic descriptors for the urban, forest, and cropland fractions, as well as reservoir
132 storage capacity to improve the representation of human activities in the basin.

133 With the above features, we expect GSHA to support hydrological model parameter estimation
134 and data-driven analysis of global streamflow as one of the most comprehensive LSH datasets
135 regarding sample size, variable dynamics, and uncertainty estimates. **Table 1** summarizes the
136 differences between GSHA and other prominent LSH datasets. Our paper is organized as follows.
137 Section 2 expands on **Table 1** and provides more details of the data included for GSHA. Section 3
138 introduces the data sources and methodologies involved in creating GSHA. Section 4 highlights the
139 key features of GSHA by conducting some analyses, followed by conclusions reached in Section 5.

140
141 **Table 1 Comparison of GSHA with other LSH datasets.** Note that we only include the CONUS
142 CAMELS dataset to represent regional LSH datasets for this comparison, as other regional CAMELS
143 share large similarity with CONUS CAMELS.

Factors	CAMELS (eg. US)	HydroATLAS	Caravan	GSIM	GSHA
Spatial extent	Regional	Global	Global	Global	Global
Sample size	671	8.5 million	6830	35002	21568
Time span	1980–2015	Static	1981–2020	1806-2016	1979-2021
Streamflow dynamics	Yes	No	Yes	Yes (statistical indices)	Yes (monthly and yearly statistical indices)
Meteorological time series	Yes	No	Yes	No	Yes
Multi data sources for meteorological variables	Yes	No	No	No	Yes (with uncertainty estimates)
Water storage dynamics	No	No	Only soil water dynamics	No	Yes
Land cover dynamics	No	No	No	No	Yes
Reservoir dynamics	No	No	No	No	Yes
Static attributes	Yes	Yes	Yes (from HydroATLAS)	Yes	Yes (from HydroATLAS)

144

2 Dataset content of GSHA v1

145 In this section, the data fields, variables, and attributes included in GSHA are described in more
146 details and summarized in **Table 2**. For the instructions of the data format, we provide a user manual
147 along with the dataset (see `readme.docx`). GSHA includes yearly and monthly streamflow
148 characteristics derived from daily discharge observations, meteorological variables (including
149 precipitation, 2-m air temperature, long- and shortwave radiation, wind speed, actual and potential
150 evapotranspiration (AET and PET)), daily or weekly water storage terms (4 layers of soil moisture,
151 groundwater, and snow depth water equivalence), daily vegetation index (leaf area index (LAI)),
152 yearly LULC characteristics (urban, cropland, and forest fraction), and yearly reservoir information
153 (degree of regulation (DOR) and reservoir capacity). For each meteorological variable, multiple
154 independent data sources are incorporated to provide uncertainty estimates. Static attributes like
155 land physiography, soils, and geology are not additionally extracted, as similar efforts have been
156 made by other researchers, so we directly matched our gauge locations to the HydroATLAS dataset
157 (Lehner et al., 2022; Linke et al., 2019) by providing the river ID match table. Users can link the
158 two to obtain these attributes.

159 **Watershed polygons:** GSHA includes 21568 watershed polygons delineated from the global
160 gauges, which are stored as Esri Shapefile format. The ID and agency of each watershed is the same
161 as the corresponding gauge ID, and the gauge latitude/longitude are in decimal degree. The area
162 denotes the upstream drainage area of the gauge. Some of the IDs contain characters (such as '.',
163 '-', etc.) inconsistent with the majority of IDs. For the convenience of the users, we unified these as
164 underscores and stored the new file names as 'filename'. We also provide independent files
165 summarizing basic information of the watersheds, including matched MERIT river reach COMID,
166 upstream area, order and downstream river reach COMID, as well as verification with officially
167 reported areas of the agencies.

168 **Streamflow indices:** GSHA publishes annual and monthly streamflow indices derived from
169 daily streamflow data, including different percentiles, and mean/median/minimum/maximum. The
170 frequency and durations of extremely high and low streamflow events are also provided. We also
171 include numbers of zero observations and valid samples to allow flexible data screening by the users.
172 The indices are stored as comma-separated values (CSV) files, with each watershed corresponding
173 to one file. A complementary R package can be used to automatically download many of the gauge
174 datasets is available at <https://github.com/Ryan-Riggs/RivRetrieve> (Riggs et al., 2023).

175 **Meteorological variables:** The meteorological variables selected are the most influential
176 drivers for streamflow, which include precipitation, 2-m temperature, ET, radiation and wind speed.
177 In main-stream land surface models, ET is a diagnostic variable derived from meteorological inputs
178 and is not considered as meteorological forcing. However, as many hydrological models also use
179 potential ET as an input variable, and model calibration sometimes involves actual ET (Immerzeel
180 & Droogers, 2008), we include the two variables and place them into the meteorological variable
181 category. For each variable, more than one data sources are used to allow for uncertainty analysis,
182 which is provided on a yearly basis in an independent file.

183 **Natural water storage terms and land use/land cover change:** These include soil moisture,
 184 snow water equivalent, and groundwater percentages. We also include yearly land cover dynamics
 185 (i.e., urban, forest, and cropland fraction changes), as well as dynamically changing reservoir
 186 capacity and degree of regulation (DOR) percentage. Leaf area index (LAI) is also included to
 187 reflect the seasonal changes in vegetation canopy that are also key to the streamflow processes.

188 **Static attributes:** GSHA does not extract updated static attributes because HydroATLAS
 189 already made substantial efforts in this regard. Instead, the listed categories are those mostly related
 190 to streamflow prediction from HydroATLAS selected to be included in GSHA files, and we direct
 191 the readers to the ID match table to access the entire 281 static attributes offered by HydroATLAS
 192 (Lehner et al., 2022; Linke et al., 2019). Our user manual, available at the dataset download site,
 193 also provides more information on it.

194

195 **Table 2 Fields provided with GSHA.**

Category	Field	Description	Unit
Watershed	Sttn_Nm	The ID of the watershed.	NaN
Polygons and basic information	Latitude	Latitude of the gauge.	Degree
	Longitude	Longitude of the gauge.	Degree
	Shedarea	The area of delineated watershed.	Km ²
	Agency	The agency the gauge belongs to.	NaN
	filename	The name of the corresponding Shapefile in the dataset.	NaN
	verification	Verification of watershed area with officially reported area of the corresponding agency. If we did not access the officially reported area of the watershed on the agency website, the field would be “unverified”.	NaN
	COMID	ID of the MERIT river reach matching with the watershed.	NaN
uparea	Upstream area of the river reach included in the MERIT database.	NaN	
order	Stream order of the river reach.	NaN	
NextDownID	ID of the downstream river reach in MERIT.	NaN	
Category	Indices	Description	Unit/Format
Streamflow indices (yearly)	percentiles	Annual 1, 10, 25, 75, 90, 99 percentiles of daily streamflow.	m ³ /s
	mean	Annual mean of daily streamflow.	m ³ /s
	median	Annual median of daily streamflow.	m ³ /s
	annual maximum flood (AMF)	Annual maximum of daily streamflow.	m ³ /s
	AMF occurrence date	The date of AMF occurrence.	Year/month/day

frequency of high-flow events	Number of days in a year with streamflow \geq 90 percentile flow.	Days/year
average duration of high-flow events	Average number of consecutive days \geq 90 percentile flow.	Days
frequency of low-flow events	Number of days in a year with streamflow \leq 10 percentile flow.	Days/year
average duration of low-flow events	Average number of consecutive days \leq 10 percentile flow.	Days
Q=0 days	Number of days with runoff=0.	Days
valid observation days	Number of days with no missing data. (Valid observations refer to non-null measurements.)	Days

Category	Indices	Description	Unit/Format
Streamflow indices (monthly)	percentiles	Monthly 1, 10, 25, 75, 90, 99 percentiles of daily streamflow.	m ³ /s
	mean	Monthly mean of daily streamflow.	m ³ /s
	median	Monthly median of daily streamflow.	m ³ /s
	monthly maximum flood (MMF)	Monthly maximum of daily streamflow.	m ³ /s
	MMF occurrence date	The date of MMF occurrence.	Year/month/day
	frequency of high-flow events	Number of days in a month with streamflow \geq yearly 90 percentile flow.	Days/month
	average duration of high-flow events	Average number of consecutive days in the month \geq yearly 90 percentile flow.	Days
	frequency of low-flow events	Number of days in a month with streamflow \leq yearly 10 percentile flow.	Days/month
	average duration of low-flow events	Average number of consecutive days in the month \leq yearly 10 percentile flow.	Days
	Q=0 days	Number of days with runoff=0.	Days
valid observation days	Number of days with no missing data.	Days	

Category	Variable	Data source name	Unit
Meteorological Variables	Precipitation	MSWEP	mm
		EM-Earth	mm
	2 m temperature	ERA5	K
		MERRA-2	K
		EUSTACE	K
	Actual	REA	mm
	evapotranspiration	GLEAM	mm
	Potential	GLEAM	mm
	evapotranspiration	hPET	mm

Radiation (longwave)	ERA5 land surface net thermal radiation	W/m ²
	MERRA-2 surface net downward longwave flux	W/m ²
Radiation (shortwave)	ERA5 land surface net solar radiation	W/m ²
	MERRA-2 surface net downward shortwave flux	W/m ²
10 m wind speed (u component)	ERA5 land u-component of wind	m/s
	MERRA-2 10 metre eastward wind	m/s
10 m wind speed (v component)	ERA5 land v-component of wind	m/s
	MERRA-2 10 metre northward wind	m/s
10 m wind speed (actual)	ERA5 land u- and v-components of wind	m/s
	MERRA-2 10 metre northward and eastward wind	m/s

Category	Variable	Data source name	Unit
Water storage terms	Soil moisture layer 1	ERA5 land soil water layer 1 (0-7 cm, 0cm refers to the surface)	m ³ /m ³
	Soil moisture layer 2	ERA5 land soil water layer 2 (7-28 cm)	m ³ /m ³
	Soil moisture layer 3	ERA5 land soil water layer 3 (28-100 cm)	m ³ /m ³
	Soil moisture layer 4	ERA5 land soil water layer 4 (100-289 cm)	m ³ /m ³
	Snow water equivalent	ERA5 land snow depth water equivalent	m of water equivalent
	Ground water	GRACE-FO data assimilation	%

Category	Variable	Data source name	Unit
Land use and land cover	Urban fraction	GAUD	%
	Forest fraction	MCD12Q1	%
	Cropland fraction	MCD12Q1	%
	Reservoir capacity	GeoDAR	Million m ³
	DOR	GeoDAR	%
	LAI	CDR LAI	NaN

Category	Attribute	Column name (directly from RiverATLAS)	Unit
Static-Physiography	Elevation	ele_mt_uav	m. a.s.l.
	Terrain slope	slp_dg_uav	degrees (x10)
	Stream gradient	sgr_dk_rav	decimetres per km
Static-Hydrology	Inundation Extent	inu_pc_ult	%
	Groundwater Table Depth	gwt_cm_cav	cm
Static-	Land Cover Classes	glc_cl_cmj	NaN

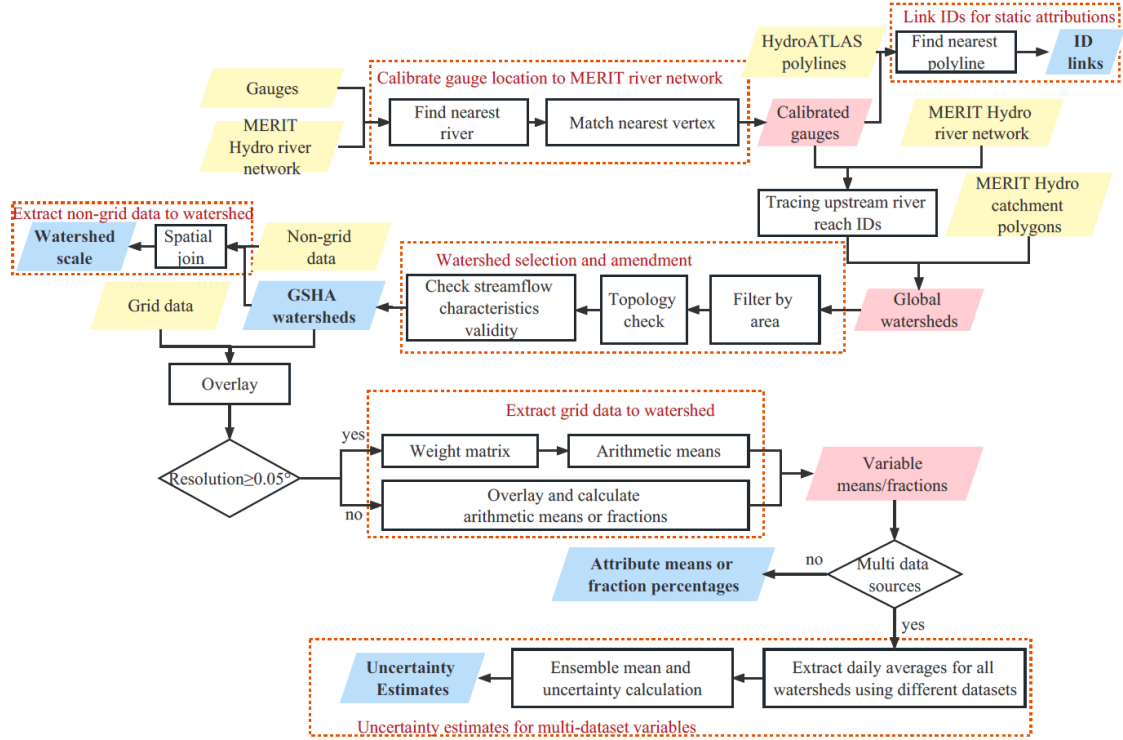
Landcover	Potential	Natural	pnv_cl_cmj	NaN
	Vegetation Classes			
	Wetland Extent		wet_pc_u01-u09	%
	Glacier Extent		gla_pc_use	%
	Permafrost Extent		prm_pc_use	%
Static-Soil	&	Clay Fraction in Soil	cly_pc_uav	%
geology		Silt Fraction in Soil	slt_pc_uav	%
		Sand Fraction in Soil	snd_pc_uav	%
		Lithological Classes	lit_cl_cmj	NaN
		Soil Erosion	ero_kh_uav	kg/hectare per year

196 3 Data sources and methodology

197 3.1 Technical workflow in creating GSHA

198 The creation of GSHA starts from revisiting the data compilation process for the stream
 199 gauging observations from 13 international agencies. The general workflow of GSHA data
 200 production processes is illustrated in **Figure 1**, which consists of watershed delineation, variable
 201 extraction from both grid and non-grid data sources, and uncertainty analysis.

202 First, we delineated the upstream watersheds using gauge locations. Calibration of gauge
 203 longitudes and latitudes were conducted to match the gauges with the MERIT river network exactly.
 204 The delineated watersheds were selected and manually checked using standards of area, topology
 205 correctness, and observation data lengths. The selected watersheds went on to be overlaid with
 206 grid and non-grid variable data sources for to obtain GSHA variables.



207
208
209
210
211

Figure 1 General workflow of GSHA. The yellow parallelograms are the input datasets, the blue ones are the final outputs of GSHA dataset, and the pink ones are the results in the process. The black quadrilaterals represent the extraction and calculation processes, and the red dotted rectangles illustrate different modules of the extraction process.

212 3.2 Gauge-based streamflow indices

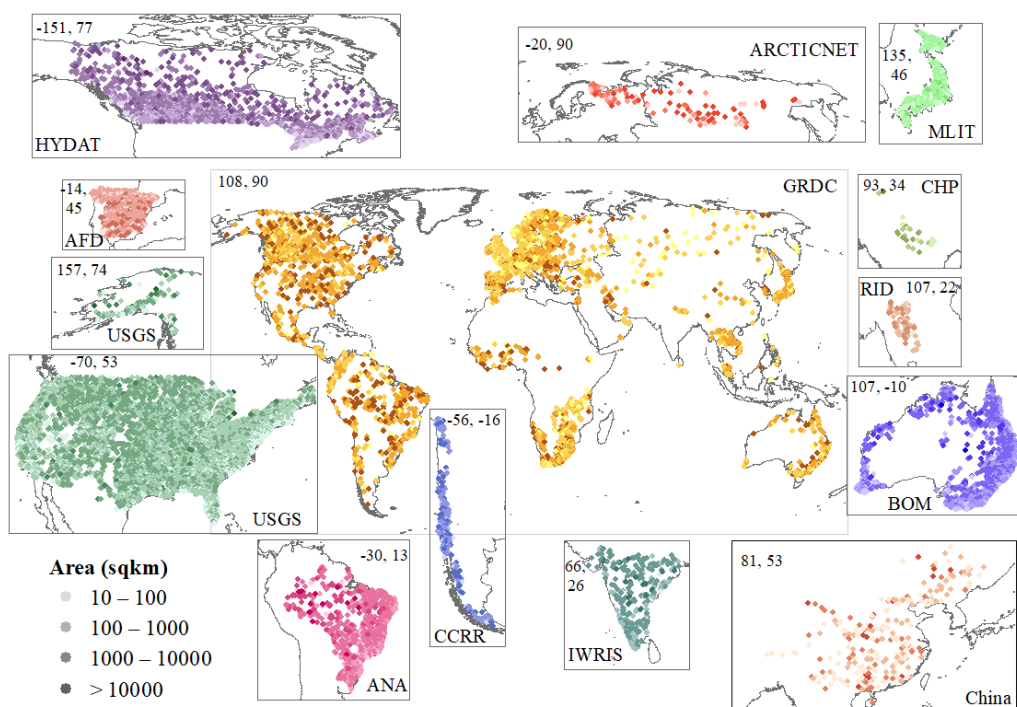
213 As shown in **Table 3**, in total streamflow data from 36497 gauges were initially scraped from
214 the web and from the Chinese National Real-time Rain and Water Situation Database. For gauges
215 located within ~100 m of each other, those with fewer years of measurements were removed,
216 assuming that they are redundant with one another. The gauge measurements were converted to a
217 consistent unit (m^3/s) and then manually compared with GRDC measurements to ensure accurate
218 unit conversion (Riggs et al., 2023). Gauge databases compiled in this study are available through
219 a variety of web interfaces, except for the Chinese Hydrology Project (CHP) data which is provided
220 by the authors of the dataset (Henck et al 2010, Schmidt et al 2011), and processed into annual scale
221 data that meets the requirements of the synthesis dataset.

Table 3 Gauge data sources used in this analysis. N1 and N2 refers to numbers of gauges with observations after 1979 and used in GSHA. The starting and ending years (Y1 and Y2) of GSHA gauges for each agency are listed.

Source	N1	N2	Y1	Y2	URL/Provider
ArcticNET 2022	116	106	1979	2003	www.r-arcnetnet.sr.unh.edu/v4.0/AllData/index.html
Australian Bureau of Meteorology 2022 (BOM)	4017	2340	1979	2021	www.bom.gov.au/waterdata/
Brazil National Water Agency 2022 (ANA)	1343	1172	1979	2021	www.snith.gov.br/hidroweb/serieshistoricas
Canada National Water Data Archive 2022 (HYDAT)	3771	2222	1979	2021	www.canada.ca/en/environment-climate-change/services/water-overview/quantity/monitoring/survey/data-products-services/national-overview/
Chile Center for Climate and Resilience Research 2022(CCCR)	481	392	1979	2020	https://explorador.ccr2.cl/
Chinese Hydrology Project (CHP)	112	26	1979	1987	(Henck et al 2010, Schmidt et al 2011)
The Global Runoff Data Centre 2022 (GRDC)	6345	4004	1979	2021	https://portal.grdc.bafg.de/applications/public.html?publicuser=PublicUser
India Water Resources Information System 2022 (IWRIS)	547	261	1979	2020	https://indiawris.gov.in/wris/#/RiverMonitoring
Japanese Water Information System 2022 (MLIT)	1023	751	1979	2019	www1.river.go.jp/
Spain Anuario de Aforos, 2022 (AFD)	1138	889	1979	2018	datos.gob.es/catalogo/e00125801-anuario-de-aforos/resource/4836b826-e7fd-4a41-950c-89b4eaea0279
Thailand Royal Irrigation Department 2022 (RID)	126	73	1980	1999	http://hydro.iis.u-tokyo.ac.jp/GAME-T/GAIN-T/routine/rid-river/disc_d.html
U.S. Geological Survey 2022 (USGS)	16951	9069	1979	2021	https://waterdata.usgs.gov/nwis/rt
Chinese National Real-time Rain and Water Situation Database	527	263	2000	2019	http://xxfb.mwr.cn/sq_zdysq.html

222 3.3 Watershed delineation

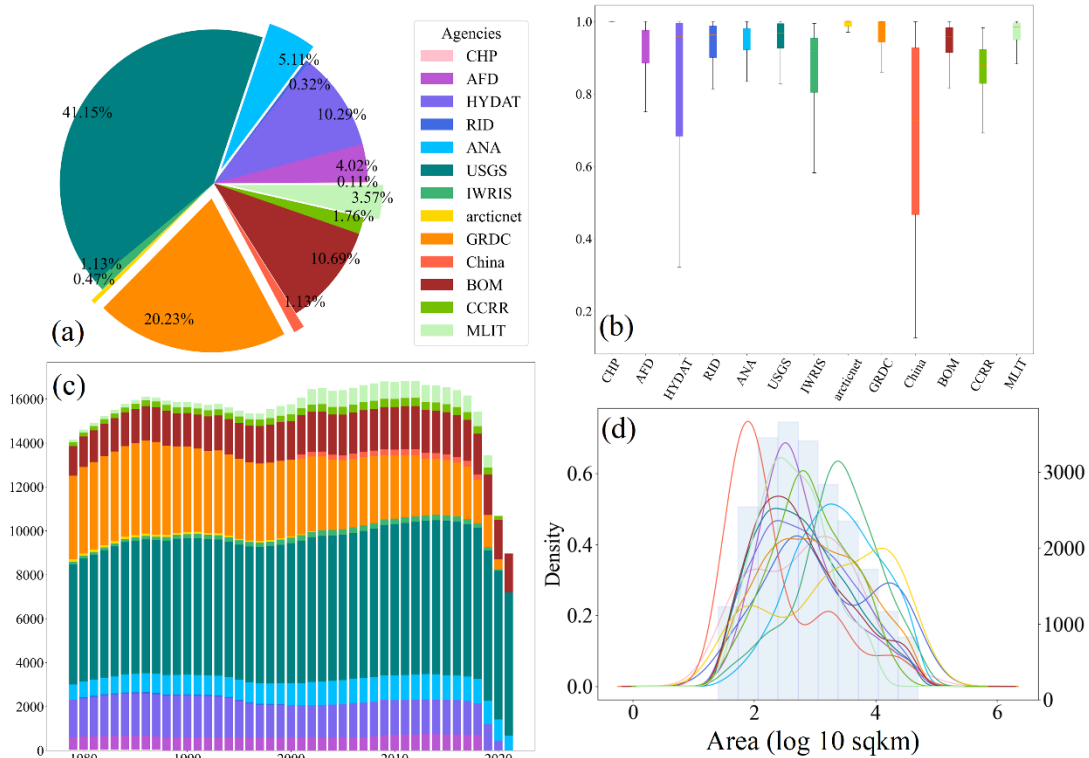
223 The watershed delineation process was built upon a vector-based global river network dataset
 224 (Lin et al., 2021), which is delineated from the 90-m Multi-Error-Removed Improved Terrain
 225 (MERIT) digital elevation model (DEM) (Yamazaki et al., 2017) and the flow direction and flow
 226 accumulation rasters (Yamazaki et al., 2019). The locations of the gauges may contain locational
 227 errors and direct delineation will result into erroneous watershed boundaries; therefore, gauge
 228 location correction was conducted by relocating the gauges to the nearest MERIT-based river reach
 229 vertices. The adjusted gauge points were used as the watershed outlets, where the contributing areas
 230 were extracted by dissolving all upstream catchments based on the topology provided by MERIT
 231 Basins (Lin et al., 2019). Since the area threshold of MERIT Basins is 25 km², we did not include
 232 watersheds smaller than this threshold. Considering the spatial heterogeneity of very large basins,
 233 we excluded watersheds $\geq 50,000$ km² from the dataset. To ensure GSHA supports studies with
 234 sufficiently long records, only watersheds with >5 years of observations since 1979 were selected.
 235 For gauges sharing the same watershed, the one with better data quality (i.e., longer measurement
 236 records and more valid observation days) was used. If the two gauges share the same quality, we
 237 only included the furthest downstream gauge. Eventually, the selection processes resulted in 21568
 238 valid watersheds out of 35970 gauges initially scraped from the web plus 527 gauges from the
 239 Chinese National Real-time Rain and Water Situation Database (**Figure 2**).



240
 241 **Figure 2 Spatial distribution of the GSHA gauges (n=21568).** Watershed areas are represented by the
 242 tint of colours. Gauges of different agencies are represented with separate colours and are plotted in
 243 individual frames (except for USGS gauges in two frames to incorporate Alaska). The agency names and
 244 the upper-left coordinates (longitude, latitude) of each frame are also shown in the figure.

245
 246 The GSHA watersheds are unevenly distributed across the globe, more than half of which are
 247 located in North America (USGS, HYDAT, and a large proportion of GRDC gauges, **Figure 3a**).

248 Europe, Australia, and South America also have relatively good coverage, while Asia and Africa
 249 show the lowest gauge densities. The majority of the gauged watersheds are of medium sizes ranging
 250 from 250 to 2500 km², although for some agencies it does not show the same distribution (**Figure**
 251 **3d**). For instance, ANA (South America), IWRIS (India), and arcticnet (Northern Eurasia)
 252 watersheds are generally larger, while the Chinese National Real-time Rain and Water Situation
 253 Database provides more gauges with smaller drainage areas. Due to the maintenance difficulties,
 254 the number of functioning gauges is declining for agencies like GRDC, but the lack of data in recent
 255 years (**Figure 3c**) is mainly due to latency issues. USGS, BOM, and ANA provide a stable number
 256 of observations for the 1980-2021 period (**Figure 3c**) with high proportions of valid observations
 257 each year (**Figure 3b**), while observational periods from arcticnet and China contain relatively fewer
 258 valid samples (**Figure 3b**) and shorter time spans (**Figure 3c**).



259
 260 **Figure 3 Summary statistics of the GSHA gauges.** This includes (a) proportions of gauges from
 261 different agencies, (b) box plots for proportions of valid observations for each agency,
 262 (c) proportion of valid observation for each year by agency and (d) distributions of watershed areas for each agency
 263 (kernel density estimation lines, left y-axis) and all gauges (blue histogram, right y-axis). The colour
 264 legend in subplot (a) applies to all four subplots. In subfigure (a) the 0.11% label corresponds to CHP,
 265 and the legend goes counter clockwise in the pie chart. In subfigure (c), CHP bars are at the bottom of
 266 the plot, and the legend goes from bottom to the top of the bars.

267 **3.4 Meteorological variables, water storage terms, and land surface characteristics**

268 After watershed delineation, publicly available grid or non-grid data were obtained and
 269 overlaid to derive the meteorological, water storage terms, and land surface characteristics. The data
 270 sources used for GSHA are listed in **Table 4**. We prioritized the use of multi-source fusion datasets

271 with relatively high quality surveyed from literature when creating GSHA.

272 3.4.1 Meteorology datasets

273 For precipitation, the Multi-Source Weighted-Ensemble Precipitation (MSWEP) that merged
274 gauge measurements (CPC Unified), grid data (GPCC), satellite products (CMORPH, GSMaP-
275 MVK, and TMPA 3B42RT), and reanalysis data (ERA-Interim and JRA-55) with sample density
276 and comparative performance considered (Beck et al., 2017; Beck et al., 2019) are included. Another
277 precipitation dataset is the Ensemble Meteorological Dataset for Planet Earth (EM-Earth)
278 deterministic estimates, which merged a station-based Serially Complete Earth (SC-Earth)
279 removing the temporal discontinuities in raw station observations and ERA5 estimates (Tang et al.,
280 2022).

281 For 2-m air temperature, the EUSTACE global land station daily air temperature dataset
282 (EUSTACE) statistically merged station and satellite observations to obtain global daily near-
283 surface air temperature (Brugnara et al., 2019) is included. Other datasets used for 2-m temperature
284 extraction are the reanalysis datasets Modern-Era Retrospective analysis for Research and
285 Applications Version 2 (MERRA-2) (Gelaro et al., 2017) and the fifth generation of European
286 Reanalysis (ERA5) dataset land component (Muñoz-Sabater et al., 2021). MERRA-2, produced
287 by NASA's Global Modelling and Assimilation Office (GMAO), used the Goddard Earth Observing
288 System (GEOS) model and analysis scheme and assimilated the latest observations. ERA5
289 reanalysis was developed by the European Centre for Medium-Range Weather Forecasts (ECMWF)
290 using the Carbon Hydrology-Tiled ECMWF Scheme for Surface Exchanges over Land
291 (CHTESSEL) driven by the downscaled meteorological forcing from the ERA5 climate reanalysis
292 (Hersbach et al., 2020). These reanalysis datasets are also used in extracting long- and shortwave
293 radiation, as well as u- and v-components of wind.

294 For AET, the REA dataset, which used the reliability ensemble averaging (REA) method to
295 merge ERA5, Global Land Data Assimilation System Version 2 (GLDAS2), and MERRA-2 is used
296 (Lu et al., 2021). Another AET data source is the product of the Global Land Evaporation
297 Amsterdam Model (GLEAM) based on satellite observations of surface net radiation and near-
298 surface air temperature (Martens et al., 2017). For PET, GLEAM is also incorporated. Another PET
299 dataset for GSHA is an hourly PET at 0.1° resolution for the global land surface (hPET) calculated
300 from ERA5-land wind speed, air and dew point temperature, net radiation components, and surface
301 air pressure (Singer et al., 2021).

302 3.4.2 Water storage term datasets

303 ERA5-land data is also applied in extracting soil moisture for 4 soil layers, as well as snow
304 water equivalence. For groundwater, an assimilation dataset from NASA's Gravity Recovery and
305 Climate Experiment (GRACE) and its follow-on mission (GRACE-FO) is used (Li et al., 2019).
306 The dataset merged water storage derived from GRACE satellite products into ECMWF Integrated
307 Forecasting System meteorological data-forced NASA's Catchment land surface model (CLSM).
308 The data is represented as groundwater drought indicator (GWI), which is the percentage of
309 groundwater storage estimates from the GRACE data assimilation relative to the climatology

310 (representing historical conditions), at weekly time scales from 2003-2021.

311 3.4.3 Land surface characteristic datasets

312 Global urban development for 1985-2015 is represented as the urban fraction in each watershed
 313 using the global annual urban dynamics (GAUD) at 30-m resolution. The dataset was derived from
 314 Landsat surface reflectance based on the Normalized Urban Areas Composite Index (NUACI) (Liu
 315 et al., 2020). For forest and cropland fractions, the Terra and Aqua combined Moderate Resolution
 316 Imaging Spectroradiometer (MODIS) Land Cover Type (MCD12Q1) land cover dataset, is used
 317 (Friedl et al., 2010). It covers 2001-2020 with a resolution of 500 m, and the categories used for
 318 GSHA are the International Geosphere–Biosphere Programme classification (IGBP) forests and
 319 croplands. Another land cover is vegetation, which is represented by LAI obtained from the National
 320 Oceanic and Atmospheric Administration (NOAA) Climate Data Record (CDR) of Advanced Very
 321 High-Resolution Radiometer (AVHRR) product, which relied on artificial neural networks and
 322 AVH09C1 surface reflectance product (Claverie et al., 2016).

323 3.4.4 Dams and reservoirs

324 The newly published Georeferenced global Dams And Reservoirs (GeoDAR) dataset that
 325 documented the dam and reservoir construction years is used for building the temporally varying
 326 watershed reservoir capacity and DOR. GeoDAR georeferenced the International Commission on
 327 Large Dams (ICOLD) World Register of Dams (WRD), and geo-matched multi-source regional
 328 registers and geocoding descriptive attributes through the Google Maps API (Wang et al., 2022).
 329 The reservoir capacities are used together with the mean annual streamflow to obtain the DOR based
 330 on equation $dor = SC/Q_{mean}$, where SC refers to reservoir storage capacity and Q_{mean} is the
 331 mean annual streamflow in the corresponding year.

332 3.4.5 Static variables

333 We matched GSHA river IDs and HydroATLAS river reach IDs to link the static attributes.
 334 HydroATLAS includes 56 variables for hydrology, physiography, climate, land cover & use, soils
 335 & geology, and anthropogenic influences for over 8.5 million river reaches globally.

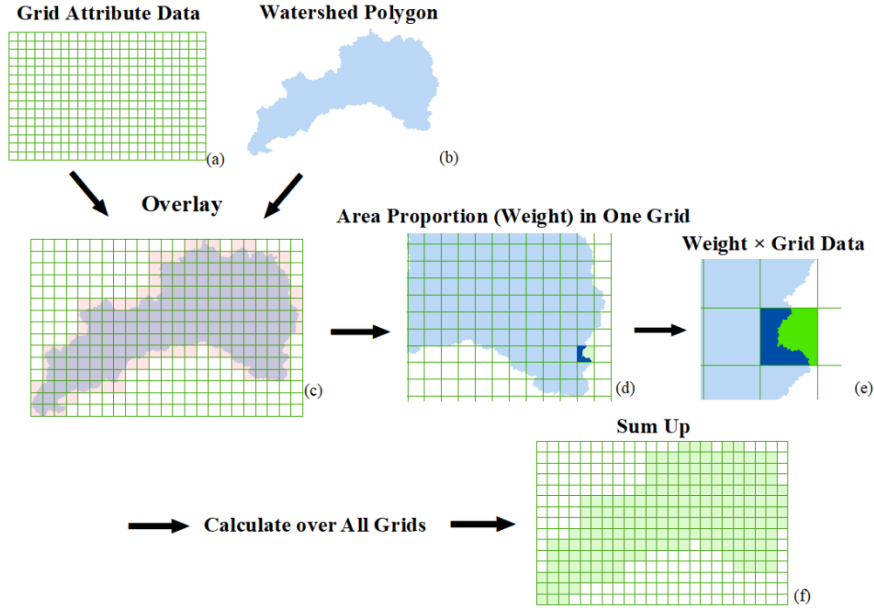
336
 337 **Table 4 Data sources used for the GSHA variables.**

Category	Dataset	Resolution	Interval	Reference
Meteorology	MSWEP	0.25°	Daily	(Beck et al., 2017; Beck et al., 2019)
	EM-Earth	0.1°	Daily	(Tang et al., 2022)
	ERA5-land	0.1°	Hourly	(Muñoz-Sabater, 2019)
	MERRA-2	0.5°* 0.625°	Hourly	(GMAO, 2015)
	EUSTACE	0.25°	Daily	(Brugnara et al., 2019)
	REA	0.25°	Daily	(Lu et al., 2021)

	GLEAM	0.25°	Daily	(Martens et al., 2017; Miralles et al., 2011)
	hPET	0.1°	Daily	(Singer et al., 2021)
Water storage terms	ERA5-land	0.1°	Hourly	(Muñoz-Sabater, 2019)
	GRACE-FO data assimilation	0.25°	Weekly	(Li et al., 2019; Zaitchik et al., 2008)
Land surface	GAUD	30 m	Yearly	(Huang, 2020)
	MCD12Q1	500 m	Yearly	(Friedl et al., 2019)
	CDR Leaf Area Index	0.05°	Daily	(Vermote et al., 2019)
Dam and reservoir	GeoDAR	NaN (polygon)	Yearly	(Wang et al., 2022)
Static Attributes	HydroATLAS	NaN (line)	NaN (static)	(Lehner et al., 2022; Linke et al., 2019)

338 3.5 Variable extraction methods

339 For grid data with relatively coarse spatial resolutions ($\geq 0.05^\circ$), we used an area-weighted
340 approach to extract the variable (Addor et al., 2017) based on the proportion of the grid area
341 contained in the basin boundary, while for high-resolution grid data, we extracted the arithmetic
342 mean directly. **Figure 4** shows the area-weighted average approach we used for grid data with spatial
343 resolution $\geq 0.05^\circ$ to reduce the influence of watershed area on data uncertainty (Tang et al., 2022).
344 The grid data (**4a**) and the quality-controlled watersheds (**4b**) were overlaid and all grids
345 intersecting with the watershed were obtained (**4c**). For each intersected grid, the proportion of the
346 polygon in the grid was calculated as the weight (dark blue, **4d**); the product of the weight and the
347 corresponding grid value was calculated over all intersected grids (**4e**) and were summed up as the
348 weighted average (**4f**). For wind, the u- and v-wind components were first used to calculate wind
349 speed, then the basin average was calculated with the weighted average approach. For grid data with
350 a spatial resolution of $< 0.05^\circ$, the area-weighted approach was not adopted as it offers limited gains
351 while becoming computationally too expensive. For reservoirs, we used the reservoir polygons in
352 GeoDAR, which were spatially joined to GSHA watershed polygons. All the intersected reservoirs
353 were considered contributory to the management of the corresponding watershed and were used to
354 calculate the total reservoir storage capacity and degree of regulation.



355

356

357

358

Figure 4 Determination of the area weights in extracting gridded data to GSHA watershed polygons. This weighted approach is applied to data at a resolution of $\geq 0.05^\circ$ but not for data at a finer spatial resolution due to computational costs.

359

3.6 Uncertainty estimates

360

361

362

363

We also provided uncertainty estimates of the meteorological variables by calculating the long-term mean of each dataset in each watershed, where the discrepancy between the maximum and minimum among the data sources (X_{max} and X_{min}) as a percentage of their mean (\bar{X}) was used in the uncertainty estimation (see Eq. 1):

364

$$uncertainty = \frac{X_{max} - X_{min}}{\bar{X}} * 100\%, \quad (1)$$

365

366

3.7 Validation

367

368

369

370

371

372

373

After delineation, we validated our watershed areas with officially reported watershed areas from BOM, HYDAT, and GRDC by matching GSHA watersheds by their agency IDs. We set the criteria of mismatched watersheds as (1) the area difference being over $\pm 20\%$ of the officially reported area, and (2) the area ratio being less than 0.1 or over 10 times the reported areas. Since not all agency websites reported watershed areas, thus we added a flag field in the attributes as “unverified”, “verified match”, and “verified mismatch” to allow users to filter the watersheds flexibly and avoid putting the samples in the dataset under an unfair standard.

374

375

376

377

Postprocessing of the extracted variables includes the unification of units and manual quality checks. For streamflow characteristics, we validated three of our indices against GSIM for its global coverage, including the mean annual streamflow, 10th and 90th percentiles. The spatial joint between GSHA and GSIM gauges in a 10 km buffer zone was performed, and only the GSIM gauge with a

378 minimum distance and watershed area difference $\leq 5\%$ to a GSHA gauge was considered. Pairs
379 with 0 measurements were excluded and 9835 pairs were involved eventually. We plotted the scatter
380 plot of GSHA-GSIM mean flow, 10th and 90th percentiles, and compared the fitting line to the 1:1
381 line, with correlation coefficients calculated (see Section 4.1).

382 We also validated precipitation, potential ET, and 2 m air temperature with the regional
383 CAMELS-US dataset. We compared the Daymet meteorological variables of CAMELS and the
384 mean of GSHA variables for validation. Since we included ERA5 data for most of our variables
385 directly or indirectly as the data source, while Caravan consistently used ERA5, we did not use
386 Caravan for the global validation as it is not considered as fully independent from GSHA. The
387 spatial match was the same as we did for GSIM which resulted in 906 pairs. This number was larger
388 than the total CAMELS gauge numbers as some gauges might be repeatedly paired due to location
389 bias of the USGS gauges and MERIT river networks, as well as the adjacency between gauges of
390 different agencies. Similarly, scatter plots and correlation coefficients are provided for assessment.

391 3.8 Watershed classification and change detection

392 We classified the watersheds as natural and human-managed to analyse the influence of human
393 water management. A watershed is classified as a natural watershed if it satisfies the following: (1)
394 DOR is smaller than 10%; (2) the urban extent is less than 5%; and (3) the sum of urban and cropland
395 fractions is smaller than 10% (L. Yang et al., 2021; Zhang et al., 2023). The classification was
396 performed for 2001-2015, and the changing patterns of the watersheds are divided into six categories:
397 (1) natural (N) when the watershed remained natural for all 15 years; (2) human managed (H) when
398 the watershed remained human managed for all 15 years; (3) natural to human managed (NH) when
399 the watershed was first natural in 2001, but changed to and maintained human managed later; and
400 (4) human managed to natural (HN) when the watershed was first human managed in 2001, but
401 changed to and maintained natural later.

402 4 Results

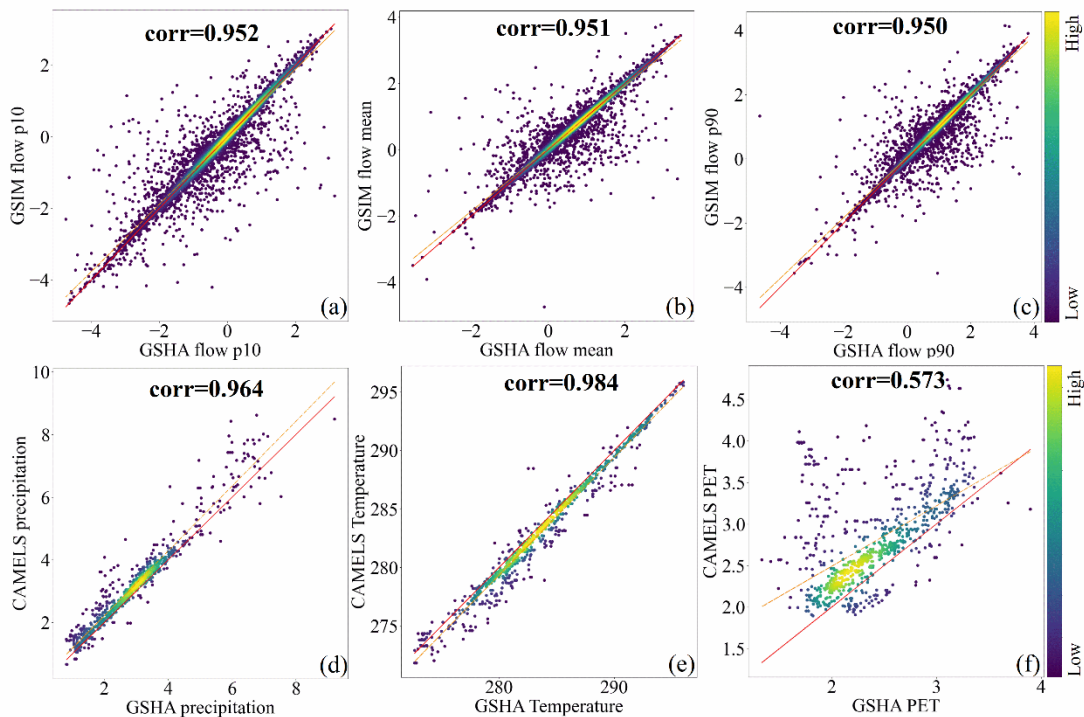
403 As previous studies have already revealed the spatial patterns of the LSH hydrometeorological
404 variables both locally and globally, here we put the spatial patterns of GSHA meteorological
405 variables and streamflow indices in **Appendix A**, while we focus on using the Results section to
406 reveal the uniqueness of GSHA. These include a technical validation of GSHA, uncertainty analysis,
407 and the temporal change of watershed human management levels.

408 4.1 Technical validation

409 The validation result figures of watershed areas are in **Appendix B** since we focused more on
410 the variables and already added the validity results in the dataset as “unverified”, “verified match”,
411 and “verified mismatch” fields in the dataset. Under our criterion of filtering “mismatch” watersheds,
412 1.9% of BOM watersheds, 4.7% of HYDAT watersheds and 8.9% of GRDC watersheds are

413 mismatched. After removing these watersheds, correlation coefficients between GSHA and the
 414 agencies can reach 0.99, which verified the correctness of our watershed delineation and data
 415 extraction approach.

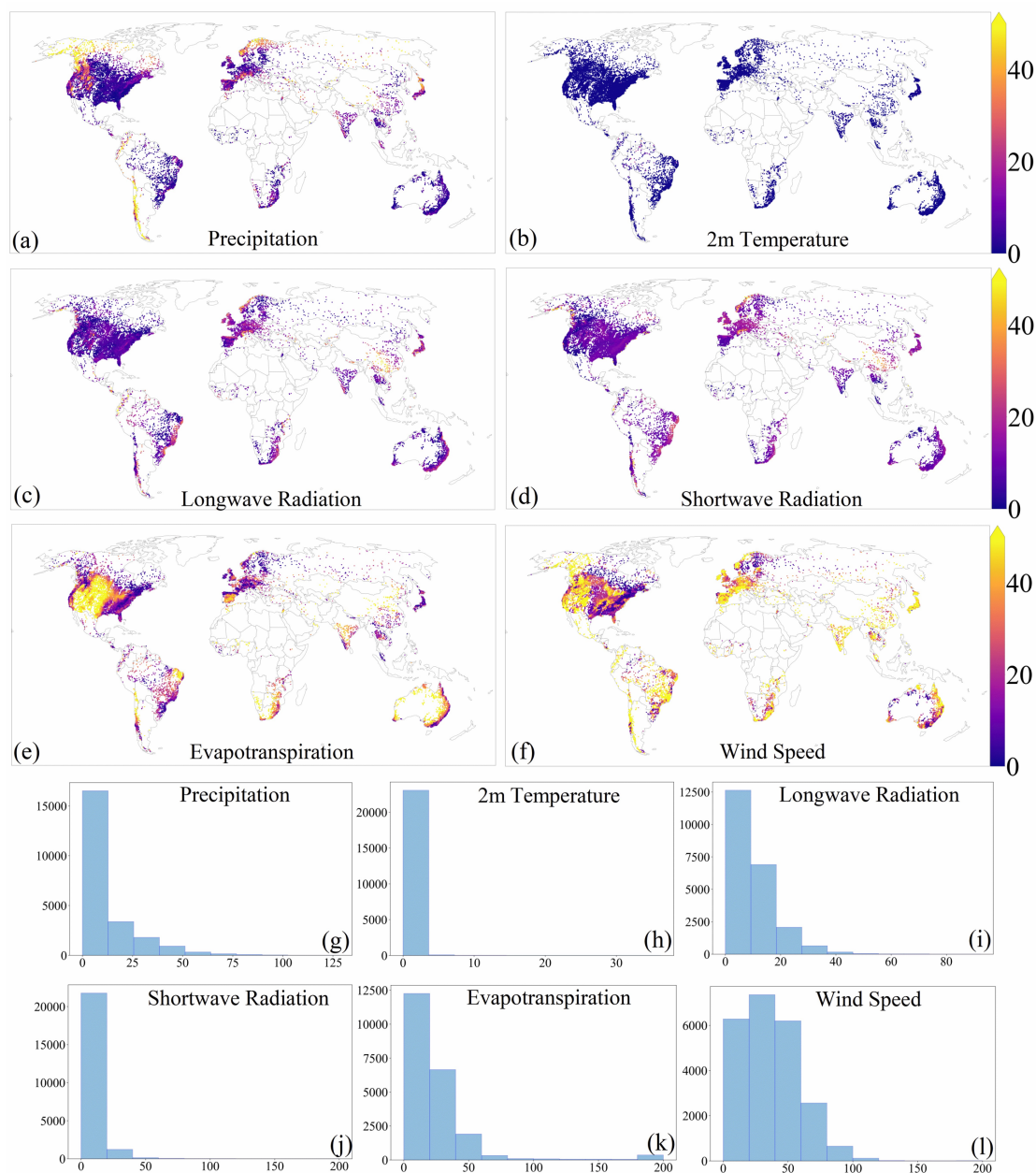
416 **Figure 5** illustrates the validation results of GSHA. **Figures 5a–5c** show streamflow indices
 417 as validated against GSIM globally, and **Figures 5d–5f** show meteorological variables as validated
 418 against Daymet from CONUS CAMELS. For streamflow indices, precipitation, and temperature,
 419 the correlation coefficients exceed 0.95 (significance $p < 0.01$), and the fitting lines are close to the
 420 1:1 line, indicating high consistencies between GSHA and the reference datasets. For PET, however,
 421 the coefficient is low, at only 0.573 (significance $p < 0.05$), and the CAMELS PET is generally higher
 422 than GSHA ensemble, which is possibly ascribed to the high uncertainty among PET datasets that
 423 is yet to be fully resolved (Singer et al., 2021) (see **Appendix C**). Note that the gauge pairing might
 424 bring a small proportion of wrong pairs for some very close gauges, and differences in temporal
 425 ranges of GSHA and GSIM might cause some discrepancies for observed streamflow.



426 **Figure 5** Validation of GSHA with GSIM streamflow characteristics ((a), (b) and (c)), and
 427 CAMELS meteorological variables ((d), (e) and (f)). ‘Corr’ in the subfigure is the Pearson correlation
 428 coefficient. The red line is the 1:1 line, while the orange dotted line is the fitting line of the scatter points.
 429 The colour bar represents density of the sample points. The unit of X and Y axes in (a), (b). and (c) is
 430 $\text{long}10 \text{ m}^3/\text{s}$.
 431

432 4.2 Uncertainty patterns for the GSHA meteorological variables

433 **Figure 6** shows the distributions of the uncertainties for different variables, and the colour bars
 434 are unified to allow for comparisons between different variables.



435

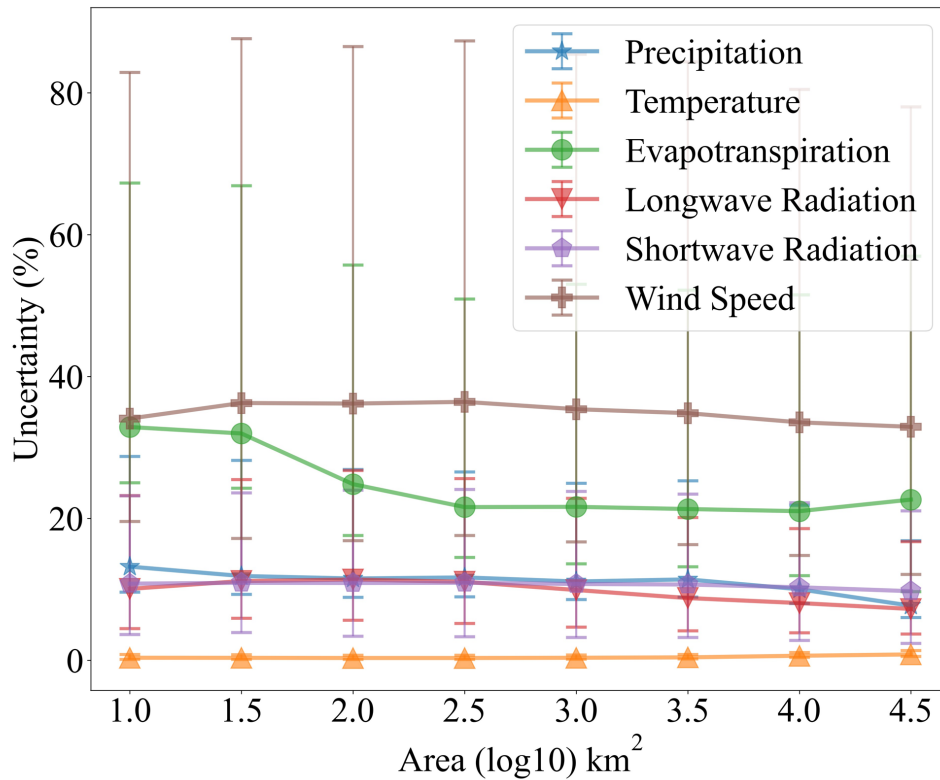
436 **Figure 6 Global patterns of the uncertainty for the GSHA meteorological variables (in percentage).**

437 This includes the uncertainty (a) for precipitation (mm/day), (b) 2-m temperature (K), (c) longwave
 438 radiation (W/m^2), (d) shortwave radiation (W/m^2), (e) evapotranspiration (mm/day), and (f) wind speed
 439 (m/s), and (g) the uncertainty histogram for precipitation, (h) 2-m temperature, (i) longwave radiation,
 440 (j) shortwave radiation, (k) evapotranspiration, and (l) wind speed.

441

442 Generally, among all variables, air temperature (**Figures 6b & 6h**) shows the minimum
 443 uncertainty (<5%), suggesting high consistency of air temperature estimates from different datasets.
 444 The uncertainty for wind speed (**Figure 6f**) is the highest among all variables. Uncertainties for
 445 other variables show strong spatial variability. For example, uncertainties for precipitation are high
 446 in high-latitude or mountainous areas like the Rocky Mountains, northern Europe, the Alps, and the
 447 Andes areas (**Figure 6a**). This is reasonable because limited accessibility to in-situ observations and
 448 the misestimation of snow (Schreiner-McGraw & Ajami, 2020) can contribute to precipitation

449 estimation errors, while the data sources show relatively high consistency (*uncertainty* $\leq 25\%$) in
450 other parts of the world (**Figure 6g**). For radiation, as solar/shortwave radiation is largely affected
451 by sky conditions, uncertainties are high in regions with less clear sky, including south-west China
452 and its surrounding areas, high latitude regions of the northern hemisphere, and Europe (Brun et al.,
453 2022). These places are also subject to high thermal/longwave radiation uncertainties for similar
454 reasons (**Figure 6c**). Land cover including vegetation and artificial surface, is another factor
455 influencing surface net radiation through the albedo effect (Hu et al., 2017), thus for heavily
456 vegetated and urbanized areas, such as the Amazon region and east coastal Australia, uncertainties
457 for both longwave and shortwave fluxes are also relatively high. Nevertheless, **Figures 6i & 6j**
458 demonstrate that for the majority of watersheds, radiation uncertainties are $< 25\%$, indicating that
459 the radiation data sources are generally consistent with each other. ET uncertainties are generally
460 larger than the above variables (**Figures 6e & 6k**), and are particularly prominent in dry areas of the
461 globe, e.g., central North America, northern Andes, central Asia, and Australia's grasslands and
462 deserts. It is also prominent in agriculture intensive regions like India and the northern part of China
463 (Sörensson & Ruscica, 2018), where agricultural irrigation may be the contributing factor to the ET
464 uncertainty. The spatial distributions of wind speed do not seem to show clear regional patterns
465 (**Figure 6f**), and uncertainty values of wind speed are generally larger over the majority of
466 watersheds (**Figure 6l**). Nevertheless, the uncertainties are low in Appalachia and northern Europe,
467 and are high in most parts of Brazil, the Andes, Africa, eastern and southern parts of Asia, as well
468 as Australia (**Figure 6f**). As we already selected relatively high-quality datasets for the variables,
469 these areas might be calling for more attention by the LSH developers, while providing possible
470 explanations for the inconsistencies in interpreting results or understanding the challenges in
471 estimating model parameters by the LSH users.



472

473 **Figure 7 Relationship between variable uncertainties and watershed areas.** The markers indicate
 474 mean values of the variable uncertainties in watersheds smaller than the corresponding x-axis value. The
 475 error bars represent the range between 25 and 75 percentiles of the uncertainty values.

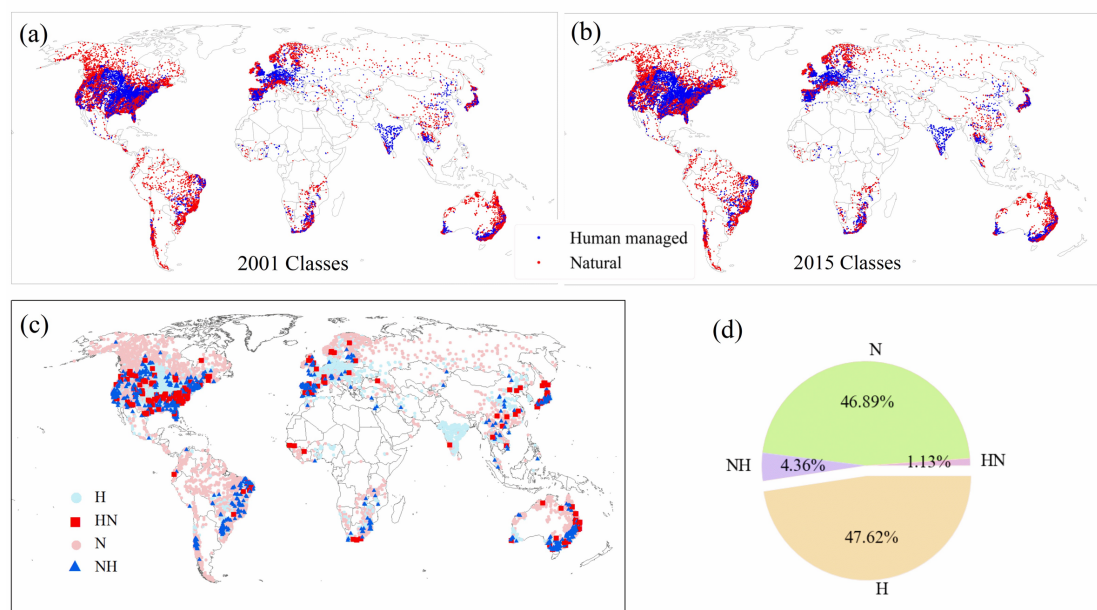
476

477 Apart from the spatial patterns above, we also investigated the emergent patterns of the
 478 uncertainties. Existing studies indicate small basins can show larger uncertainties due to coarse
 479 resolution data inputs (Kauffeldt et al., 2013), while sub-grid variabilities might be offset by
 480 averaging over large watersheds. As we plotted the uncertainty against watershed areas in **Figure 7**,
 481 it verifies that for most variables, the uncertainty declines as the watershed area increases. **Figure 7**
 482 also reveals some interesting patterns which were rarely discussed in existing studies. For example,
 483 the most obvious decline of data uncertainty with area came from ET (green). ET is highly
 484 dependent on and significantly affected by land surface spatial heterogeneity, thus it benefits the
 485 most from spatial averaging for large river basins. Longwave radiation uncertainty (red) experiences
 486 a moderate decline, likely due to its linkage with land surface complexity and cloud conditions.
 487 Shortwave radiation and precipitation uncertainty show a similar decline pattern (blue and purple),
 488 which is possibly related to their strong ties to cloud covers. Temperature has a low uncertainty, and
 489 its relationship to watershed area is also not obvious. Wind speed uncertainty only declines slightly
 490 as the area increases, and this may be because wind speed uncertainty can be traced back more to
 491 the atmospheric circulation patterns instead of land surface conditions, thus showing a non-
 492 prominent relationship with watershed area. Overall, GSHA provides uncertainty estimates that
 493 capture these prominent patterns, which can be helpful to hydrologic modellers and users.

494 4.3 Natural and human managed watersheds and changing patterns

495 We also demonstrate the other key features of GSHA by categorizing global watersheds into
 496 natural and human-managed, and more prominently their temporal shifts in **Figure 8**. Overall, the
 497 majority of human-managed watersheds are located in the US, Europe, and other regions with
 498 intensive industrial or agricultural activities such as East and South Asia (**Figures 8a and 8b**).
 499 During 2001-2015, 46.89% of the watersheds remained natural, while another 47.62% under human
 500 management in 2001 remained in the category throughout the study period (**Figure 8d**). Generally,
 501 the northern hemisphere has a larger proportion of human-managed watersheds, while watersheds
 502 in the less populated and urbanized southern hemisphere largely remain natural.

503 Noticeably, 4.36% of GSHA watersheds switched from natural to human-managed (1011
 504 watersheds), and the remaining 1.13% changed back to natural states from human managed during
 505 2001-2015. For instance, watersheds in the middle and lower Yangtze River area and the north-
 506 eastern China show a shift from human-managed to natural state, where ecological restoration
 507 projects were in place (Qu et al., 2018; Zhang et al., 2015). Although the time span of GSHA LULC
 508 dynamics restricted the change detection for developed countries as their urbanizations and
 509 infrastructure developments have long been completed, and for fast emerging economies after 2015,
 510 the time series were also missing; nevertheless, the changing human activities captured by GSHA
 511 may be helpful to understand the streamflow changes including flood characteristics (Yang et al.,
 512 2021; Zhang et al., 2022).



513

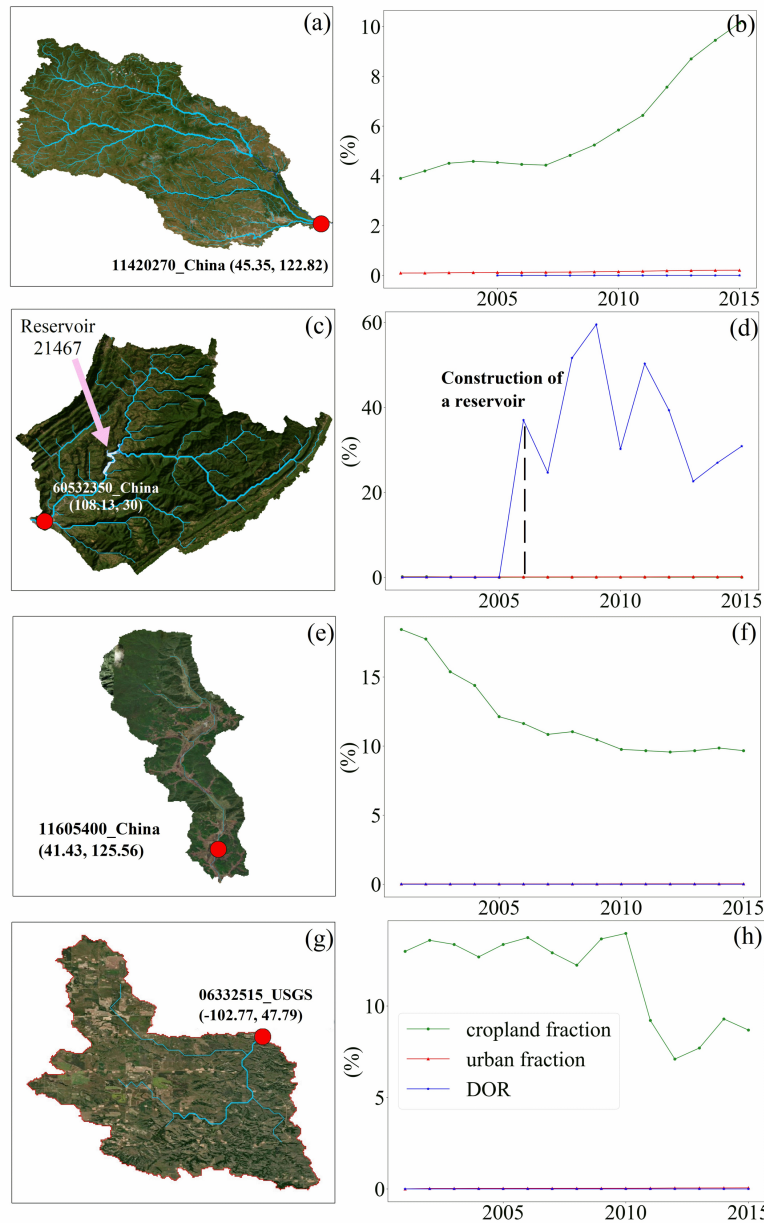
514 **Figure 8 Classification of natural and human managed watersheds in 2001 (a) and 2015 (b).**

515 **Changes in watershed categories are illustrated by (c) and (d).** H and N in (c) and (d) represent
 516 watersheds that maintained human managed or natural from 2001-2015; NH and HN represent those
 517 changing from natural to human managed and from human managed to natural, respectively.

518

519 We further used several examples to illustrate the changing status of GSHA watersheds (**Figure**
 520 **9**). **Figures 9a and 9b** show a watershed located in Northeast China, where the rapid increase in

521 cropland shifted the watershed from natural states to human-managed in recent years. **Figures 9c**
 522 **and 9d** correspond to a mountainous area in Sichuan Province, China, which became human-
 523 managed due to the construction of a reservoir in 2006. For another case in Northeast China
 524 (**Figures 9e and 9f**) and a USGS case (**Figures 9g and 9h**), the watersheds shifted from human-
 525 managed to natural, which is mainly manifested by the reduction in cropland fraction due to the
 526 environmental policy. For instance, afforestation during 2000-2010 in Changbai Mountains where
 527 the watershed in **Figures 9e and 9f** is located, significantly increased the forest cover and might
 528 bring a decline in human disturbance in the form of land use (Zhang & Liang, 2014). These results
 529 highlight the shifting watershed status that would require further attention from LSH users, which
 530 is encapsulated in GSHA v1.0 and will be continuously improved in the future.



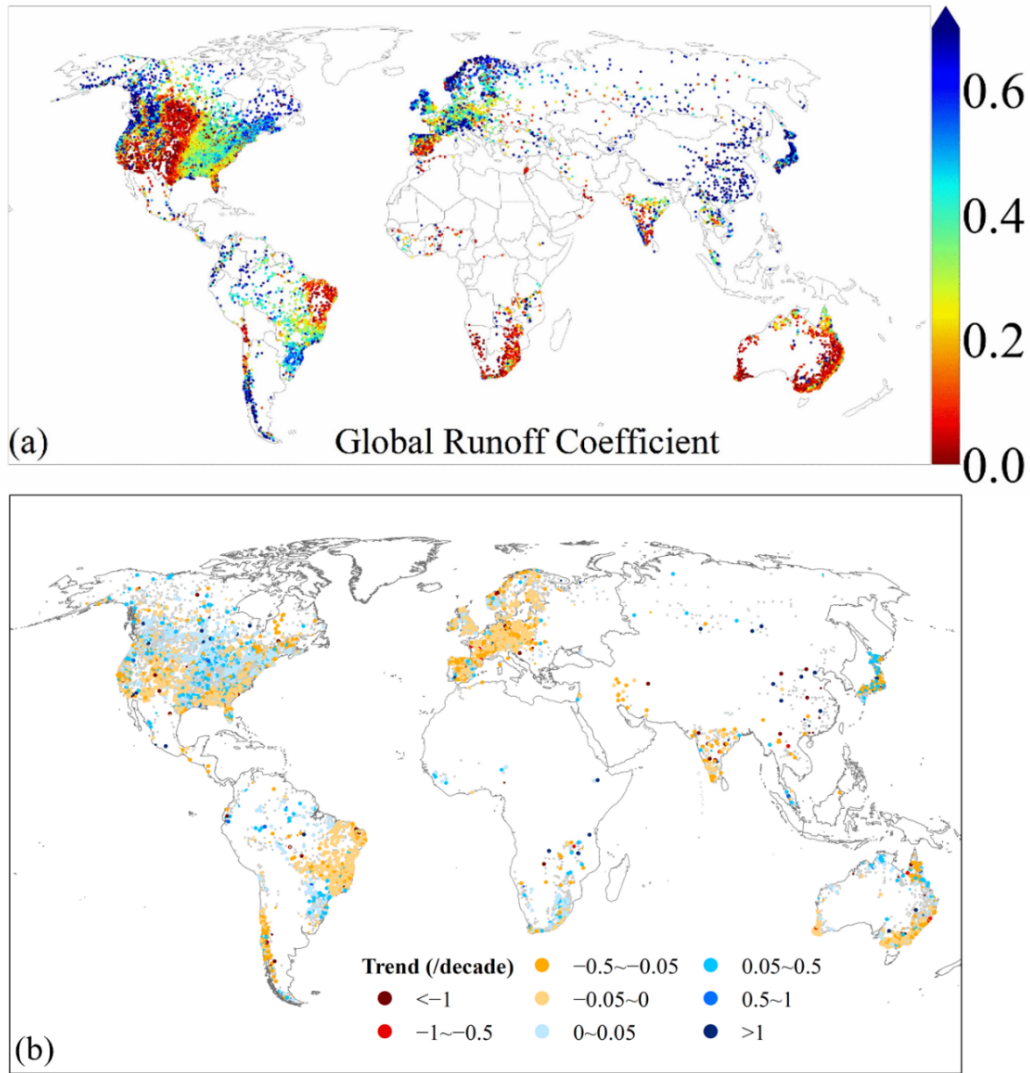
531 **Figure 9** Cases for shifting status of the watershed classification. (a) and (b) correspond to
 532 11420270_China, and (c) and (d) correspond to 60532350_China, both of which changed from natural
 533 to human managed category. (e) and (f) represent 11605400_China, and (g) and (h) correspond to
 534

535 06332515_USGS watershed changing from human managed to natural watershed.

536 4.4 Changing runoff coefficient patterns derived from GSHA

537 Finally, we also analysed the global pattern in the trend of runoff coefficient (RC) as a brief
538 demonstration on what GSHA can offer out of its many potential usages. RC is defined as R/P ,
539 where R denotes runoff (mm) and P denotes precipitation (mm). **Figure 10a** shows that regions with
540 high RC (i.e., a large proportion of rainfall goes into rivers instead of being evaporated or consumed)
541 are in east Asia and North America, most parts of Europe, the west coast of North America and the
542 Amazon, in general agreement with the aridity patterns across the globe. For arid/semiarid areas
543 and places with intense water use (e.g., western US, eastern Brazil, Australia, Africa), RC is low,
544 meaning most of the precipitation does not reach the gauged river.

545 We found that RC generally remained stable for the past decades (i.e., grey dots in **Figure**
546 **10b**; >80% of the gauges did not observe a statistically significant trend), while 4252 watersheds
547 observed a statistically significant trend in RC at 95% level (5690 watersheds at 90% level). Among
548 them, decreasing RC is more widespread than increasing RC. The most pronounced decreasing
549 trends are observed in Europe, India, eastern Brazil, Chile, eastern Australia, and the Euphrates and
550 Tigris, which largely correspond to regions with known intense agricultural, industrial, and
551 residential water use that may have reduced the river water. We note that the global RC trend patterns
552 were different from a recent study that showed mostly increasing RC in the high-latitudes, central
553 North America, eastern Australia, and Europe (Xiong et al., 2022). Given Xiong et al. (2022) used
554 estimated runoff while we used runoff directly from gauge observations, it is likely that the
555 concerning water availability issues in the context of increasing human water use may not be fully
556 captured by existing studies. Regional studies also tend to show inconsistent results. For example,
557 a study based on models incorporating climate change and land use change but ignoring human
558 water consumptions suggested that deforestation and urbanization generally increase RC (Lucas-
559 Borja et al., 2020), while another study identified a significant decreasing trend for RC by focusing
560 on cases with intense irrigational water use (Banasik and Hejduk, 2012). These collectively preclude
561 a clear identification of consistent RC trends (Velpuri and Senay, 2013) and a clear causal factor
562 attribution given the complexity of the anthropogenic factors. As such, GSHA may offer a new path
563 to fill in the gap of disentangling the influences of large-scale water use on decreasing RC.



564

565

566

567

568

569

570

571

Figure 10 Patterns of runoff coefficient (a) and its trend (b). Only watersheds with statistically significant trend ($p<0.05$) are shown with colours in (b); the small and large sized points represent 95% ($p<0.05$) and 90% significance level ($p<0.1$), respectively. Note that the temporal coverage is different for different gauges; readers can refer to the GSHA temporal coverage for interpreting the patterns. The figure illustrates 18987 GSHA watersheds. Watersheds with less than 10 years of indices calculated from over 250 valid observations per year, as well as with runoff coefficient trend over 20 per decade, are not demonstrated in subfigure b.

572

5 Conclusions

573

574

575

576

577

Large sample hydrology (LSH) datasets play a critical role in data-driven analyses and model parameter estimation for hydrological studies. From MOPEX (Duan et al., 2006) to Caravan (Kratzert et al., 2023), significant efforts have been made to improve the comprehensiveness of LSH, yet issues related to data spatial coverage, uncertainty estimates, and human activity dynamics remain to be solved. This study complements existing LSH with a new synthesis dataset named the

578 Global Streamflow characteristics, Hydrometeorology, and catchment Attributes for large sample
579 river-centric studies (GSHA v1.1).

580 To summarize, GSHA contributes the following aspects to the LSH development:

- 581 1. It includes streamflow indices, hydrometeorological data, and surface characteristics data for
582 21568 gauges compiled from 13 agencies worldwide, which represents one of the most
583 comprehensive LSH by far.
- 584 2. We incorporated multiple data sources to provide uncertainty estimates for each meteorological
585 variable (including precipitation, 2 m air temperature, radiation, wind, and ET). The spatial
586 patterns and the relationship between the uncertainty and the watershed characteristics GSHA
587 reveals may be helpful to identify inconsistencies among data-driven studies or biases for model
588 parameter estimation studies using existing LSH.
- 589 3. Dynamic data are provided for previously static data descriptors for land cover changes
590 including urban, cropland and forest fractions, as well as reservoir storage change including
591 storage capacity and degree of regulation.

592 Although GSHA does not cover watersheds of $<25\text{km}^2$ or the dynamics of cryosphere variables
593 (e.g., glacier and permafrost) that have become increasingly important in terrestrial hydrological
594 changes, and the time spans for the dynamic descriptors of LULC are unable to cover the critical
595 periods for the advanced and less-advanced economies due to the constraints with existing LULC
596 data, GSHA is expected to be utilized to unravel the following insights:

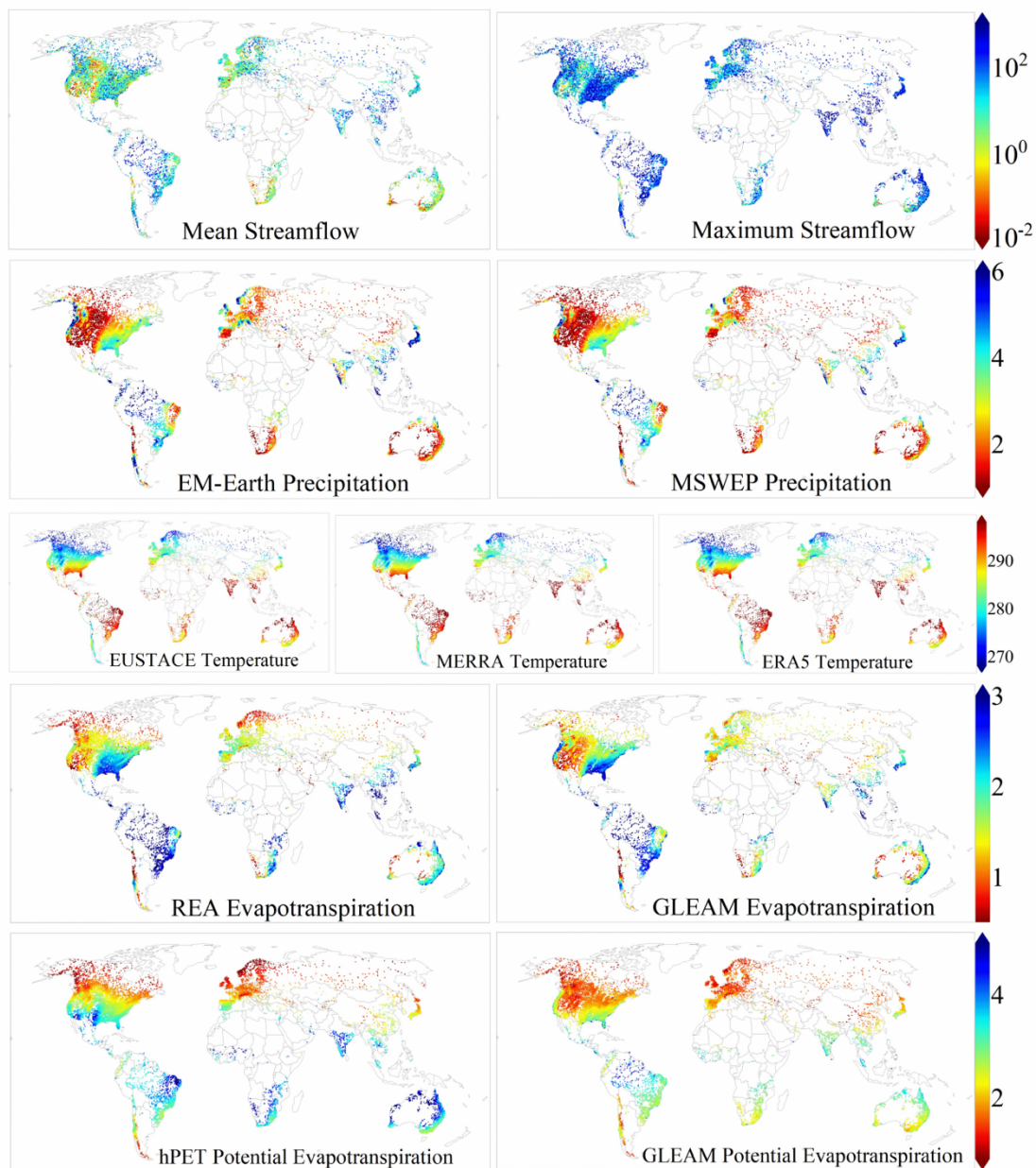
- 597 1. The uncertainty patterns vary between variables and geographical regions, indicating that the
598 interpretation of model and analysis results need to consider inconsistencies of raw data, apart
599 from looking into the methodologies and patterns themselves.
- 600 2. Although most watersheds have remained natural or human managed throughout the GSHA
601 time span, a considerable number of watersheds shifted between the two categories, which can
602 be ascribed to urbanization, cropland increase, reservoir construction and ecological restoration
603 such as returning farmland to natural states, and these can be clearly manifested using GSHA.
- 604 3. Analysis with runoff coefficient reveals that among gauges with a statistically significant trend,
605 a greater portion experienced a declining RC trend than an increase trend. This pattern revealed
606 by GSHA can be used to further study water availability issues in a changing climate.

607 As our knowledge on the above processes continues to improve, we expect that future versions
608 of GSHA will be continuously updated. Finally, better hydrological data sharing is crucial to
609 advance global change hydrology studies.

610 Appendix

611 A. Spatial patterns of GSHA meteorological variables

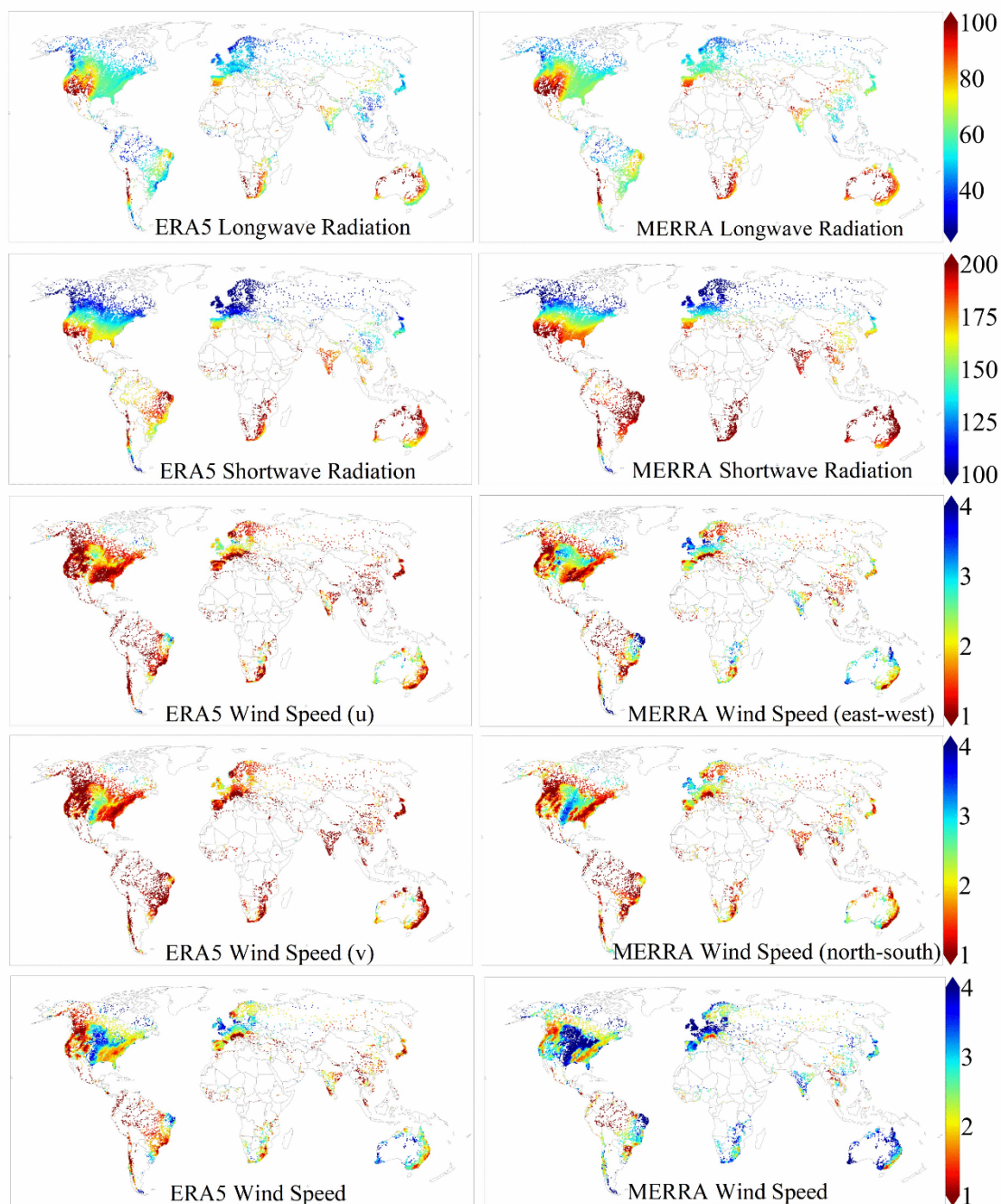
612 **Figures A1 & A2** show the spatial distributions of GSHA meteorological variables and selected
613 streamflow indices. The spatial pattern derived from each individual data source is plotted separately.



614

615 **Figure A1** Spatial distribution of streamflow indices (row 1, m^3/s), precipitation (row 2, mm/day), 2 m

616 air temperature (row 3, K), actual ET (row 4, mm/day), potential ET (row 5, mm/day).



617

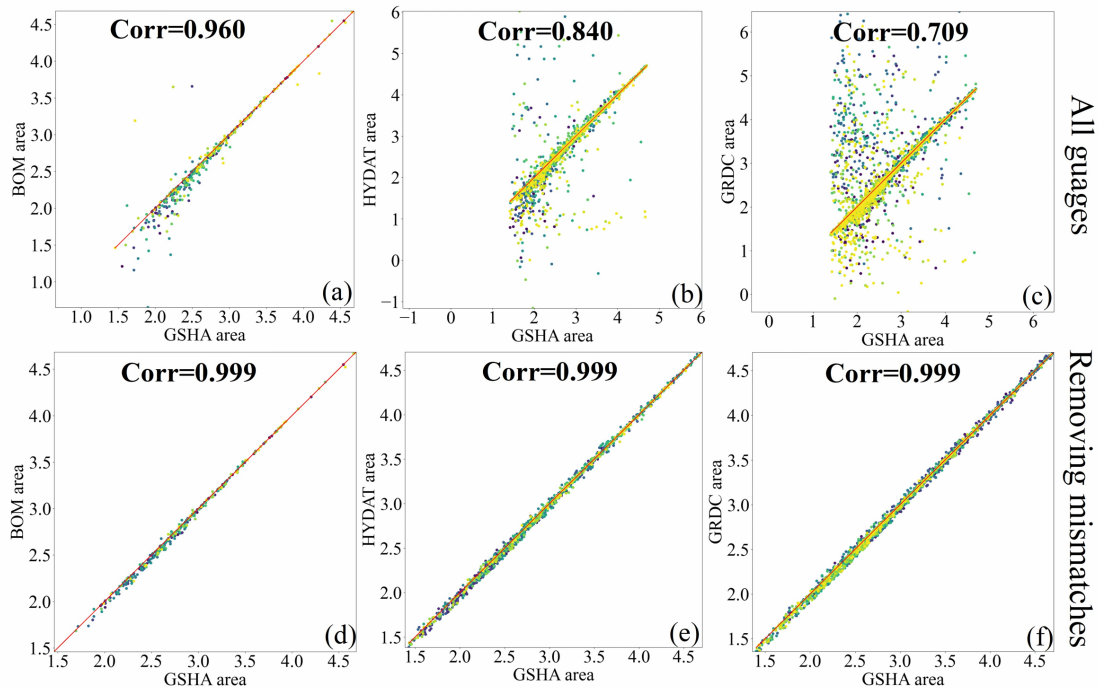
618 **Figure A 2** Spatial distribution of longwave radiation (row 1, W/m^2), shortwave radiation (row 2, W/m^2),
 619 wind u- (row 3, m/s) and v- components (row 4, m/s) and the wind speed (row 5, m/s).

620

621 **B. Validation results of watershed areas**

622 The validation results with BOM, HYDAT, and GRDC on watershed areas are plotted in
 623 **Figure B1** and **B2**, where the mismatches between GSHA areas and the officially reported areas are
 624 shown. Before removing the mismatched watersheds, their correlation coefficients are 0.960, 0.840,
 625 0.709, respectively, as shown in **Figure B1 (a), (b), (c)**. After removing the mismatched watersheds,
 626 correlation coefficients for all three agencies reach 0.999, as shown in **Figure B1 (d), (e), (f)**. As we

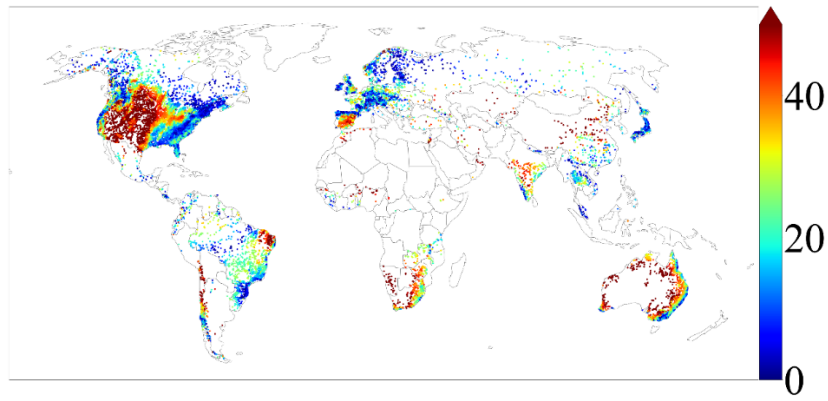
627 traced the MERIT Basins (Lin et al., 2019) for our watershed delineation, the mismatches are
 628 believed to occur when the gauge locates in the vicinity of the intersection point of a river reach and
 629 its main stream, which makes it difficult to decide which reach the gauge belongs to while matching
 630 the gauge to the MERIT river network. This explains why in **Figure B1** most of the mismatches
 631 appear at relatively small areas. As we do not have access to all official watershed areas, and **Figure**
 632 **B1 (a), (b), (c)** suggest that matching qualities differ among the agencies, to simply remove the
 633 mismatched watersheds or to modify them might put the samples in the dataset under an unfair
 634 standard. Additionally, some agencies such as GRDC experienced some updates of their gauge
 635 locations and upstream areas, thus watershed boundaries in all datasets mentioned might come with
 636 uncertainties. Therefore, we gave the watersheds as “unverified”, “verified match”, and “verified
 637 mismatch” identifiers to allow users to flexibly filter the watersheds.



638
 639 **Figure B1** Validation of GSHA with officially reported areas of BOM (a, d), HYDAT (b, e), and GRDC
 640 (c, f). Subfigures (a) to (c) are the results before removing the mismatched watersheds, and subfigures
 641 (d) to (f) represent results after removing the mismatched watersheds. The Pearson correlation coefficient
 642 are represented by “Corr” in the figure. The areas are represented by the unit of (log10 km²).

643 C. Potential evapotranspiration uncertainty

644 The spatial and numerical distributions of potential evapotranspiration (PET) uncertainties are
 645 illustrated in **Figure C1** and **Figure C2**. PET uncertainty is high compared with other variables (see
 646 5.2 section). The majority of high PET uncertainty watersheds are in dry areas, but since it is
 647 calculated from meteorological variables, exceptions exist for palces including eastern Pacific coast,
 648 where the climate is dry but PET uncertainty is low, and India, which is located in a wet climate
 649 zone but has high PET uncertainty. As demonstrated by **Figure C3**, PET uncertainty do not decrease
 650 with the increase of watershed area, probably because PET is calculated from various variables, and
 651 the calculation over large watersheds involves more uncertainties for individual grids.

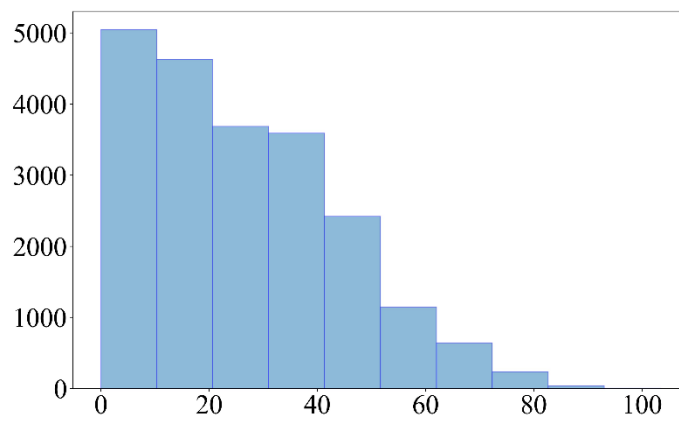


652

653

654

Figure C1 Spatial pattern of potential evapotranspiration (PET) uncertainty.

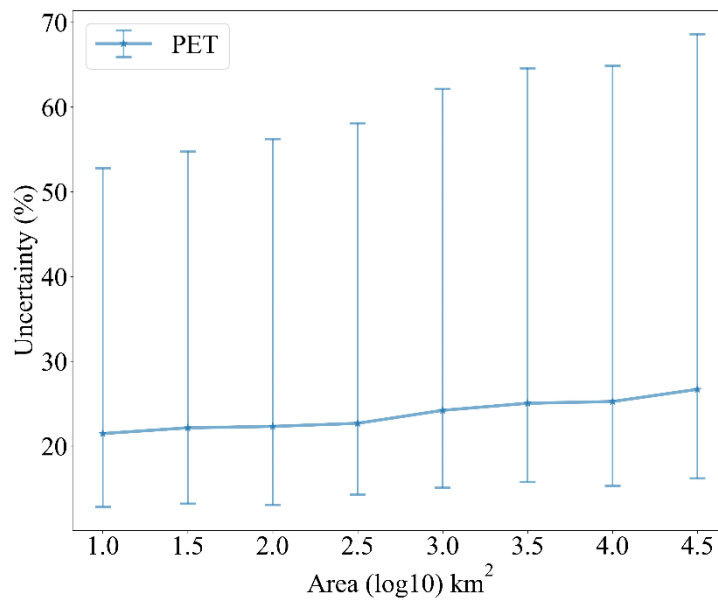


655

656

657

Figure C2 Numerical distribution of PET uncertainty.



658

659

Figure C3 Relationship of PET uncertainty to watershed area.

660 Author contribution

661 Conceptualization: PL. Investigation: ZY, PL, RR, GA, XL. Data curation: ZY, RR, XL, PL, ZZ,
662 SC. Funding acquisition: PL. Writing - initial: ZY, PL. Writing - Review and Editing: PL, ZY, GA,
663 RR, XL.

664 Data and Code Availability

665 GSHA v1.0 is openly available at <https://doi.org/10.5281/zenodo.8090704> and
666 <https://doi.org/10.5281/zenodo.10127757>. The codes involved in the workflow to generating GSHA
667 will be available upon reasonable requests to the corresponding author.

668 Competing interests

669 The authors declare no conflict of interest.

670 Acknowledgements

671 This study is supported by the National Key Research and Development Program
672 (2022YFF0801303), the Yunnan Provincial Basic Research Project-Science and Technology Special
673 Project of Southwest United Graduate School (#305107035054), the Natural Science Foundation of
674 China (42371481, 42175178), and the Fundamental Research Funds for the Central Universities to
675 Peking University (#7100604136).

676 References

- 677 Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., & Mendoza, P. A. (2020). Large-
678 sample hydrology: recent progress, guidelines for new datasets and grand challenges.
679 *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, *65*(5), 712-725.
680 <https://doi.org/10.1080/02626667.2019.1683182>
- 681 Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., & Clark, M. P. (2018). A ranking of
682 hydrological signatures based on their predictability in space. *Water Resources Research*,
683 *54*, 8792–8812. <https://doi.org/10.1029/2018WR022606>
- 684 Addor, N., & Clark, M. P. (2017). The CAMELS data set: catchment attributes and meteorology for
685 large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293-5313.
686 <https://doi.org/10.5194/hess-21-5293-2017>
- 687 Aerts, J. P., Hut, R. W., van de Giesen, N. C., Drost, N., van Verseveld, W. J., Weerts, A. H., & Hazenberg,

- 688 P. (2022). Large-sample assessment of varying spatial resolution on the streamflow estimates of
 689 the wflow_sbm hydrological model. *Hydrology and Earth System Sciences*, 26(16), 4407-4430 ,
 690 <https://doi.org/10.5194/hess-26-4407-2022>.
- 691 AghaKouchak, A., Chiang, F., Huning, L. S., Love, C. A., Mallakpour, I., Mazdiyasn, O., Moftakhari,
 692 H., Papalexiou, S. M., Ragno, E., & Sadegh, M. (2020). Climate Extremes and Compound
 693 Hazards in a Warming World. *Annual Review of Earth and Planetary Sciences*, Vol 48, 2020,
 694 48, 519-548. <https://doi.org/10.1146/annurev-earth-071719-055228>
- 695 Alvarez-Garretón, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini,
 696 M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., & Ayala, A. (2018). The
 697 CAMELS-CL dataset: catchment attributes and meteorology for large sample studies - Chile
 698 dataset. *Hydrology and Earth System Sciences*, 22(11), 5817-5846.
 699 <https://doi.org/10.5194/hess-22-5817-2018>
- 700 ArcticNET 2022 ArcticNET V1.0 (available at: <https://russia-arcticnet.sr.unh.edu/>)
- 701 Arsenault, R., Brissette, F., Martel, J.-L., Troin, M., Lévesque, G., Davidson-Chaput, J., Gonzalez, M.
 702 C., Ameli, A., & Poulin, A. (2020). A comprehensive, multisource database for
 703 hydrometeorological modeling of 14,425 North American watersheds. *Scientific Data*, 7(1), 243.
 704 <https://doi.org/10.1038/s41597-020-00583-2>
- 705 Australian Bureau of Meteorology 2022 Australian Bureau of Meteorology (available at:
 706 www.bom.gov.au/)
- 707 Banasik, K., and Hejduk, L. "Long-term changes in runoff from a small agricultural catchment." *Soil and*
 708 *Water Research* 7, no. 2 (2012): 64-72. <https://doi.org/10.1007/s10661-006-0769-2>
- 709 Beck, H. E., van Dijk, A. I., De Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., & Bruijnzeel,
 710 L. A. (2016). Global-scale regionalization of hydrologic model parameters. *Water Resources*
 711 *Research*, 52(5), 3599-3622. <https://doi.org/10.1002/2015WR018247>
- 712 Beck, H. E., Van Dijk, A. I., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., & De Roo, A.
 713 (2017). MSWEP: 3-hourly 0.25 global gridded precipitation (1979–2015) by merging gauge,
 714 satellite, and reanalysis data. *Hydrology and Earth System Sciences*, 21(1), 589-615.
 715 <https://doi.org/10.5194/hess-21-589-2017>
- 716 Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Van Dijk, A. I., McVicar, T. R., & Adler,
 717 R. F. (2019). MSWEP V2 global 3-hourly 0.1 precipitation: methodology and quantitative
 718 assessment. *Bulletin of the American Meteorological Society*, 100(3), 473-500.
 719 <https://doi.org/10.1175/BAMS-D-17-0138.1>
- 720 Belvederesi, C., Zaghloul, M. S., Achari, G., Gupta, A., & Hassan, Q. K. (2022). Modelling river flow in
 721 cold and ungauged regions: A review of the purposes, methods, and challenges. *Environmental*
 722 *Reviews*, 30(1), 159-173. <https://doi.org/10.1139/er-2021-0043>
- 723 Benke, K. K., Lowell, K. E., & Hamilton, A. J. (2008). Parameter uncertainty, sensitivity analysis and
 724 prediction error in a water-balance hydrological model. *Mathematical and Computer Modelling*,
 725 47(11-12), 1134-1149. <https://doi.org/10.1016/j.mcm.2007.05.017>
- 726 Beven, K. J., & Alcock, R. E. (2012). Modelling everything everywhere: a new approach to decision-
 727 making for water management under uncertainty. *Freshwater Biology*, 57, 124-132.
 728 <https://doi.org/10.1111/j.1365-2427.2011.02592.x>
- 729 Bourdin, D. R., Fleming, S. W., & Stull, R. B. (2012). Streamflow modelling: a primer on applications,
 730 approaches and challenges. *Atmosphere-Ocean*, 50(4), 507-536.
 731 <https://doi.org/10.1080/07055900.2012.734276>

- 732 Brazil National Water Agency 2022 National water and sanitation agency (ANA) Agência Nac Águas E
 733 Saneam. Básico ANA (available at: www.gov.br/ana/en/national_water_agency)
- 734 Brugnara, Y., Good, E., Squintu, A. A., van der Schrier, G., & Brönnimann, S. (2019). The EUSTACE
 735 global land station daily air temperature dataset. *Geoscience Data Journal*, *6*(2), 189-204.
 736 <https://doi.org/10.1002/gdj3.81>
- 737 Brun, P., Zimmermann, N. E., Hari, C., Pellissier, L., & Karger, D. N. (2022). Global climate-related
 738 predictors at kilometer resolution for the past and future. *Earth System Science Data*, *14*(12),
 739 5573-5603. <https://doi.org/10.5194/essd-14-5573-2022>
- 740 Brunner, M. I., Slater, L., Tallaksen, L. M., & Clark, M. (2021). Challenges in modeling and predicting
 741 floods and droughts: A review. *Wiley Interdisciplinary Reviews: Water*, *8*(3), e1520.
 742 <https://doi.org/10.1002/wat2.1520>
- 743 Burges, S. J. (1998). Streamflow prediction: capabilities, opportunities, and challenges. *Hydrologic
 744 Sciences: Taking Stock and Looking Ahead*, *5*, 101-134.
- 745 Canada National Water Data Archive 2022 National water data archive HYDAT (available at:
 746 [www.canada.ca/en/environment-climate-change/services/water-
 747 overview/quantity/monitoring/survey/data-products-services/national-archive-hydat.html](http://www.canada.ca/en/environment-climate-change/services/water-overview/quantity/monitoring/survey/data-products-services/national-archive-hydat.html))
- 748 Chagas, V. B., Chaffe, P. L., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C., & Siqueira, V. A.
 749 (2020). CAMELS-BR: hydrometeorological time series and landscape attributes for 897
 750 catchments in Brazil. *Earth System Science Data*, *12*(3), 2075-2096.
 751 <https://doi.org/10.5194/essd-12-2075-2020>
- 752 Chen, X., Jiang, L., Luo, Y., and Liu, J.: A global streamflow indices time series dataset for large-sample
 753 hydrological analyses on streamflow regime (until 2022), *Earth Syst. Sci. Data*, *15*, 4463–4479,
 754 <https://doi.org/10.5194/essd-15-4463-2023>
- 755 Chile Center for Climate and Resilience Research 2022 Center for climate and resilience research CR2 |
 756 Chilean research center on climate, climate change and resilience (available at: www.cr2.cl/eng/)
- 757 Cho, K., & Kim, Y. (2022). Improving streamflow prediction in the WRF-Hydro model with LSTM
 758 networks. *Journal of Hydrology*, *605*, 127297. <https://doi.org/10.1016/j.jhydrol.2021.127297>
- 759 Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J., Tang, G., Gharari, S., Freer,
 760 J. E., Whitfield, P. H., & Shook, K. R. (2021). The abuse of popular performance metrics in
 761 hydrologic modeling. *Water Resources Research*, *57*(9), e2020WR029001.
 762 <https://doi.org/10.1029/2020WR029001>
- 763 Claverie, M., Matthews, J. L., Vermote, E. F., & Justice, C. O. (2016). A 30+ year AVHRR LAI and
 764 FAPAR climate data record: Algorithm description and validation. *Remote Sensing*, *8*(3), 263.
 765 <https://doi.org/10.3390/rs8030263>
- 766 Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J., Lane, R., Lewis,
 767 M., & Robinson, E. L. (2020). CAMELS-GB: hydrometeorological time series and landscape
 768 attributes for 671 catchments in Great Britain. *Earth System Science Data*, *12*(4), 2459-2483.
 769 <https://doi.org/10.5194/essd-12-2459-2020>
- 770 Olivier Delaigue, Pierre Brigode, Vazken Andréassian, Charles Perrin, Pierre Etchevers, et al..
 771 CAMELS-FR: A large sample hydroclimatic dataset for France to explore hydrological
 772 diversity and support model benchmarking. *IAHS-2022 Scientific Assembly*, May 2022,
 773 Montpellier, France. hal-03687235
- 774 Do, H. X., Gudmundsson, L., Leonard, M., & Westra, S. (2018). The Global Streamflow Indices and
 775 Metadata Archive (GSIM) - Part 1: The production of a daily streamflow archive and metadata.

- 776 *Earth System Science Data*, 10(2). <https://doi.org/10.5194/essd-10-765-2018>
- 777 Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H., Gusev, Y., Habets, F., Hall, A.,
778 & Hay, L. (2006). Model Parameter Estimation Experiment (MOPEX): An overview of science
779 strategy and major results from the second and third workshops. *Journal of Hydrology*, 320(1-
780 2), 3-17. <https://doi.org/10.1016/j.jhydrol.2005.07.031>
- 781 Fang, Y., Huang, Y., Qu, B., Zhang, X., Zhang, T., & Xia, D. (2022). Estimating the Routing Parameter
782 of the Xin'anjiang Hydrological Model Based on Remote Sensing Data and Machine Learning.
783 *Remote Sensing*, 14(18), 4609. <https://doi.org/10.3390/rs14184609>
- 784 Fowler, K. J. A., Acharya, S. C., Addor, N., Chou, C. C., & Peel, M. C. (2021). CAMELS-AUS:
785 hydrometeorological time series and landscape attributes for 222 catchments in Australia. *Earth*
786 *System Science Data*, 13(8), 3847-3867. <https://doi.org/10.5194/essd-13-3847-2021>
- 787 Friedl, M. A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., & Huang, X. (2010).
788 MODIS Collection 5 global land cover: Algorithm refinements and characterization of new
789 datasets. *Remote Sensing of Environment*, 114(1), 168-182.
790 <https://doi.org/10.1016/J.RSE.2009.08.016>
- 791 Friedl, M., D. Sulla-Menashe. MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m
792 SIN Grid V006. 2019, distributed by NASA EOSDIS Land Processes DAAC,
793 <https://doi.org/10.5067/MODIS/MCD12Q1.006>.
- 794 Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov,
795 A., Bosilovich, M. G., & Reichle, R. (2017). The modern-era retrospective analysis for research
796 and applications, version 2 (MERRA-2). *Journal of Climate*, 30(14), 5419-5454.
797 <https://doi.org/10.1175/JCLI-D-16-0758.1>
- 798 Global Modeling and Assimilation Office (GMAO) (2015), inst3_3d_asm_Cp: MERRA-2 3D IAU State,
799 Meteorology Instantaneous 3-hourly (p-coord, 0.625x0.5L42), version 5.12.4, Greenbelt, MD,
800 USA: Goddard Space Flight Center Distributed Active Archive Center (GSFC DAAC),
801 <https://doi.org/10.5067/VJAFPLI1CSIV>.
- 802 Gudmundsson, L., Do, H. X., Leonard, M., & Westra, S. (2018). The Global Streamflow Indices and
803 Metadata Archive (GSIM) - Part 2: Quality control, time-series indices and homogeneity
804 assessment. *Earth System Science Data*, 10(2). <https://doi.org/10.5194/essd-10-787-2018>
- 805 Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., & Andréassian, V. (2014).
806 Large-sample hydrology: a need to balance depth with breadth. *Hydrology and Earth System*
807 *Sciences*, 18(2), 463-477. <https://doi.org/10.5194/hess-18-463-2014>
- 808 Hao, Z., Jin, J., Xia, R., Tian, S., Yang, W., Liu, Q., Zhu, M., Ma, T., Jing, C., & Zhang, Y. (2021). CCAM:
809 China catchment attributes and meteorology dataset. *Earth System Science Data*, 13(12), 5591-
810 5616. <https://doi.org/10.5194/essd-13-5591-2021>
- 811 Henck, A. C., Montgomery, D. R., Huntington, K. W., & Liang, C. (2010). Monsoon control of effective
812 discharge, Yunnan and Tibet. *Geology*, 38(11), 975-978. <https://doi.org/10.1130/G31444.1>
- 813 Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, Á., Muñoz-Sabater, J., Nicolas, J. P., Peubey,
814 C., Radu, R., Schepers, D., Simmons, A. J., Soci, C., Abdalla, S., Abellan, X., Balsamo, G.,
815 Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., . . . Thépaut, J. (2020). The ERA5 global
816 reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049.
817 <https://doi.org/10.1002/qj.3803>
- 818 Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J.,
819 Sivapalan, M., Pomeroy, J., Arheimer, B., Blume, T., Clark, M., & Ehret, U. (2013). A decade
of Predictions in Ungauged Basins (PUB)—a review. *Hydrological sciences journal*, 58(6),

- 820 1198-1255. <https://doi.org/10.1080/02626667.2013.803183>
- 821 Hu, D., Cao, S., Chen, S., Deng, L., & Feng, N. (2017). Monitoring spatial patterns and changes of
822 surface net radiation in urban and suburban areas using satellite remote-sensing data.
823 *International Journal of Remote Sensing*, 38(4), 1043-1061.
824 <https://doi.org/10.1080/01431161.2016.1275875>
- 825 Huang, Yinghuai (2020): High spatiotemporal resolution mapping of global urban change from 1985 to
826 2015. figshare. Dataset. <https://doi.org/10.6084/m9.figshare.11513178.v1>
- 827 Immerzeel, W., and, & Droogers, P. (2008). Calibration of a distributed hydrological model based on
828 satellite evapotranspiration. *Journal of Hydrology*, 349(3-4), 411-424.
829 <https://doi.org/10.1016/j.jhydrol.2007.11.017>
- 830 India Water Resources Information System 2022 India Water Resources Information System (available
831 at: <https://indiawris.gov.in/wris/#/>)
- 832 Japanese Water Information System 2022 Ministry of Land, Infrastructure, Transport and Tourism
833 (available at: www.mlit.go.jp/en/)
- 834 Kauffeldt, A., Halldin, S., Rodhe, A., Xu, C.-Y., & Westerberg, I. K. (2013). Disinformative data in large-
835 scale hydrological modelling. *Hydrology and Earth System Sciences*, 17(7), 2845-2857.
836 <https://doi.org/10.5194/hess-17-2845-2013>
- 837 Klingler, C., Schulz, K., & Herrnegger, M. (2021). LamaH-CE: LARge-SaMple DATA for Hydrology and
838 Environmental Sciences for Central Europe. *Earth System Science Data*, 13(9), 4529-4565.
839 <https://doi.org/10.5194/essd-13-4529-2021>
- 840 Kovács, G. (1984). Proposal to construct a coordinating matrix for comparative hydrology. *Hydrological
841 sciences journal*, 29(4), 435-443. <https://doi.org/10.1080/02626668409490961>
- 842 Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019a). Benchmarking
843 a catchment-aware long short-term memory network (LSTM) for large-scale hydrological
844 modeling. *Hydrol. Earth Syst. Sci. Discuss*, 2019, 1-32. <https://doi.org/10.5194/hess-2019-368>
845
- 846 Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019b). Towards
847 learning universal, regional, and local hydrological behaviors via machine learning applied to
848 large-sample datasets. *Hydrology and Earth System Sciences*, 23(12), 5089-5110.
849 <https://doi.org/10.5194/hess-23-5089-2019>, 2019.
- 850 Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A.,
851 Klotz, D., & Nevo, S. (2023). Caravan-A global community dataset for large-sample hydrology.
852 *Scientific Data*, 10(1), 61. <https://doi.org/10.1038/s41597-023-01975-w>
- 853 Lehner, B., Messenger, M. L., Korver, M. C., & Linke, S. (2022). Global hydro-environmental lake
854 characteristics at high spatial resolution. *Scientific Data*, 9(1), 351.
855 <https://doi.org/10.1038/s41597-022-01425-z>
- 856 Li, B., Rodell, M., Kumar, S., Beaudoin, H. K., Getirana, A., Zaitchik, B. F., de Goncalves, L. G.,
857 Cossetin, C., Bhanja, S., & Mukherjee, A. (2019). Global GRACE data assimilation for
858 groundwater and drought monitoring: Advances and challenges. *Water Resources Research*,
859 55(9), 7564-7586. <https://doi.org/10.1029/2018WR024618>
- 860 Lin, P., Rajib, M. A., Yang, Z. L., Somos-Valenzuela, M., Merwade, V., Maidment, D. R., Wang, Y., &
861 Chen, L. (2018). Spatiotemporal evaluation of simulated evapotranspiration and streamflow
862 over Texas using the WRF-Hydro-RAPID modeling framework. *JAWRA Journal of the
863 American Water Resources Association*, 54(1), 40-54. <https://doi.org/10.1111/1752-1688.12585>

- 864 Lin, P. R., Pan, M., Beck, H. E., Yang, Y., Yamazaki, D., Frasson, R., David, C. H., Durand, M., Pavelsky,
865 T. M., Allen, G. H., Gleason, C. J., & Wood, E. F. (2019). Global Reconstruction of Naturalized
866 River Flows at 2.94 Million Reaches. *Water Resources Research*, *55*(8), 6499-6516.
867 <https://doi.org/10.1029/2019wr025287>
- 868 Lin, P. R., Pan, M., Wood, E. F., Yamazaki, D., & Allen, G. H. (2021). A new vector-based global river
869 network dataset accounting for variable drainage density. *Scientific Data*, *8*(1).
870 <https://doi.org/ARTN 2810.1038/s41597-021-00819-9>
- 871 Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine,
872 V., Maxwell, S., & Moidu, H. (2019). Global hydro-environmental sub-basin and river reach
873 characteristics at high spatial resolution. *Scientific Data*, *6*(1), 283.
874 <https://doi.org/10.1038/s41597-019-0300-6>
- 875 Liu, X., Huang, Y., Xu, X., Li, X., Li, X., Ciais, P., Lin, P., Gong, K., Ziegler, A. D., & Chen, A. (2020).
876 High-spatiotemporal-resolution mapping of global urban change from 1985 to 2015. *Nature*
877 *Sustainability*, *3*(7), 564-570. <https://doi.org/10.1038/s41893-020-0521-x>
- 878 Lu, J., Wang, G., Chen, T., Li, S., Hagan, D. F. T., Kattel, G., Peng, J., Jiang, T., & Su, B. (2021). A
879 harmonized global land evaporation dataset from model-based products covering 1980–2017.
880 *Earth System Science Data*, *13*(12), 5879-5898. <https://doi.org/10.5194/essd-13-5879-2021>
- 881 Lucas-Borja, M. E., Carrà, B. G., Nunes, J. P., Bernard-Jannin, L., Zema, D. A., Zimbone, S. M., (2020)
882 Impacts of land-use and climate changes on surface runoff in a tropical forest watershed (Brazil),
883 *Hydrological Sciences Journal*, *65*:11, 1956-1973, <https://doi.org/10.1016/j.catena.2006.04.015>
- 884 Martens, B., Miralles, D. G., Lievens, H., Van Der Schalie, R., De Jeu, R. A., Fernández-Prieto, D., Beck,
885 H. E., Dorigo, W. A., & Verhoest, N. E. (2017). GLEAM v3: Satellite-based land evaporation
886 and root-zone soil moisture. *Geoscientific Model Development*, *10*(5), 1903-1925.
887 <https://doi.org/10.5194/gmd-10-1903-2017>
- 888 Merchant, C. J., Paul, F., Popp, T., Ablain, M., Bontemps, S., Defourny, P., Hollmann, R., Lavergne, T.,
889 Laeng, A., & De Leeuw, G. (2017). Uncertainty information in climate data records from Earth
890 observation. *Earth System Science Data*, *9*(2), 511-527. [https://doi.org/10.5194/essd-9-511-](https://doi.org/10.5194/essd-9-511-2017)
891 [2017](https://doi.org/10.5194/essd-9-511-2017)
- 892 Miralles, D. G., Holmes, T., De Jeu, R., Gash, J., Meesters, A., & Dolman, A. (2011). Global land-surface
893 evaporation estimated from satellite-based observations. *Hydrology and Earth System Sciences*,
894 *15*(2), 453-469. <https://doi.org/10.5194/hess-15-453-2011>
- 895 Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S.,
896 Choulga, M., Harrigan, S., & Hersbach, H. (2021). ERA5-Land: A state-of-the-art global
897 reanalysis dataset for land applications. *Earth System Science Data*, *13*(9), 4349-4383.
898 <https://doi.org/10.5194/essd-13-4349-2021>
- 899 Muñoz Sabater, J., (2019): ERA5-Land hourly data from 1981 to present. Copernicus Climate Change
900 Service (C3S) Climate Data Store (CDS), 10.24381/cds.e2161bac
- 901 Nandi, S., & Reddy, M. J. (2022). An integrated approach to streamflow estimation and flood inundation
902 mapping using VIC, RAPID and LISFLOOD-FP. *Journal of Hydrology*, *610*, 127842.
903 <https://doi.org/10.1016/j.jhydrol.2022.127842>
- 904 Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D.,
905 Brekke, L., Arnold, J. R., Hopson, T., & Duan, Q. (2015). Development of a large-sample
906 watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics
907 and assessment of regional variability in hydrologic model performance. *Hydrology and Earth*

- 908 *System Sciences*, 19(1), 209-223. <https://doi.org/10.5194/hess-19-209-2015>
- 909 Niraula, R., Meixner, T., & Norman, L. M. (2015). Determining the importance of model calibration for
910 forecasting absolute/relative changes in streamflow from LULC and climate changes. *Journal*
911 *of Hydrology*, 522, 439-451. <https://doi.org/10.1016/j.jhydrol.2015.01.007>
- 912 Qu, S., Wang, L., Lin, A., Zhu, H., & Yuan, M. (2018). What drives the vegetation restoration in Yangtze
913 River basin, China: climate change or anthropogenic factors? *Ecological Indicators*, 90, 438-
914 450. <https://doi.org/10.1016/j.ecolind.2018.03.029>
- 915 Razavi, T., & Coulibaly, P. (2013). Streamflow prediction in ungauged basins: review of regionalization
916 methods. *Journal of hydrologic engineering*, 18(8), 958-975.
917 [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0000690](https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690)
- 918 Ren, K., Fang, W., Qu, J., Zhang, X., & Shi, X. (2020). Comparison of eight filter-based feature selection
919 methods for monthly streamflow forecasting—three case studies on CAMELS data sets. *Journal*
920 *of Hydrology*, 586, 124897. <https://doi.org/10.1016/j.jhydrol.2020.124897>
- 921 Riggs, R. M., Allen, G. H., Wang, J., Pavelsky, T. M., Gleason, C. J., David, C. H., & Durand, M. (2023).
922 Extending global river gauge records using satellite observations. *Environmental Research*
923 *Letters*. <https://doi.org/10.1088/1748-9326/acd407>
- 924 Schaake, J., Cong, S., and Duan, Q. 2006. "U.S. MOPEX DATA SET". United States.
925 <https://www.osti.gov/servlets/purl/899413>. Schmidt, A. H., Montgomery, D. R., Huntington, K.
926 W., & Liang, C. (2011). The question of communist land degradation: new evidence from local
927 erosion and basin-wide sediment yield in Southwest China and Southeast Tibet. *Annals of the*
928 *Association of American Geographers*, 101(3), 477-496.
929 <https://doi.org/10.1080/00045608.2011.560059>
- 930 Schreiner-McGraw, A. P., & Ajami, H. (2020). Impact of uncertainty in precipitation forcing data sets
931 on the hydrologic budget of an integrated hydrologic model in mountainous terrain. *Water*
932 *Resources Research*, 56(12), e2020WR027639. <https://doi.org/10.1029/2020WR027639>
- 933 Singer, M. B., Asfaw, D. T., Rosolem, R., Cuthbert, M. O., Miralles, D. G., MacLeod, D., Quichimbo, E.
934 A., & Michaelides, K. (2021). Hourly potential evapotranspiration at 0.1 resolution for the
935 global land surface from 1981-present. *Scientific Data*, 8(1), 224.
936 <https://doi.org/10.1038/s41597-021-01003-9>
- 937 Spain Anuario de Aforos 2022 Anuario de Aforos—Anuario de Aforos Digital—datos.gob.es (available
938 at: [http://datos.gob.es/es/catalogo/e00125801-anuario-de-aforos/resource/4836b826-e7fd-
939 4a41-950c-89b4eaca0279](http://datos.gob.es/es/catalogo/e00125801-anuario-de-aforos/resource/4836b826-e7fd-4a41-950c-89b4eaca0279))
- 940 Sörensson, A. A., & Ruscica, R. C. (2018). Intercomparison and uncertainty assessment of nine
941 evapotranspiration estimates over South America. *Water Resources Research*, 54(4), 2891-2908.
942 <https://doi.org/10.1002/2017WR021682>
- 943 Tang, G., Clark, M. P., & Papalexiou, S. M. (2022). EM-Earth: The ensemble meteorological dataset for
944 planet Earth. *Bulletin of the American Meteorological Society*, 103(4), E996-E1018.
945 <https://doi.org/10.1175/BAMS-D-21-0106.1>
- 946 Tang, G., Clark, M., Papalexiou, S. (2022) EM-Earth: The Ensemble Meteorological Dataset for Planet
947 Earth. Federated Research Data Repository. <https://doi.org/10.20383/102.0547>
- 948 Tang, G., Clark, M. P., Knoben, W. J. M., Liu, H., Gharari, S., Arnal, L., et al. (2023). The impact of
949 meteorological forcing uncertainty on hydrological modeling: A global analysis of cryosphere
950 basins. *Water Resources Research*, 59, e2022WR033767. <https://doi.org/10.1029/2022WR033767>
- 951 Thackeray, C. W., Hall, A., Norris, J., & Chen, D. (2022). Constraining the increased frequency of global

- 952 precipitation extremes under warming. *Nature Climate Change*, 12(5), 441-448.
 953 <https://doi.org/10.1038/s41558-022-01329-1>
- 954 Thailand Royal Irrigation Department 2022 RID River Discharge Data (available at: [http://hydro.iis.u-](http://hydro.iis.u-tokyo.ac.jp/GAME-T/GAIN-T/routine/rid-river/disc_d.html)
 955 [tokyo.ac.jp/GAME-T/GAIN-T/routine/rid-river/disc_d.html](http://hydro.iis.u-tokyo.ac.jp/GAME-T/GAIN-T/routine/rid-river/disc_d.html))
- 956 The Global Runoff Data Centre 2022 The global runoff data centre GRDC Data Portal (available at:
 957 <https://portal.grdc.bafg.de/applications/public.html?publicuser=PublicUser>)
- 958 Ukhurebor, K. E., Azi, S. O., Aigbe, U. O., Onyancha, R. B., & Emegha, J. O. (2020). Analyzing the
 959 uncertainties between reanalysis meteorological data and ground measured meteorological data.
 960 *Measurement*, 165, 108110. <https://doi.org/10.1016/j.measurement.2020.108110>
- 961 U.S. Geological Survey 2019 Gages Through the Ages (available at:
 962 <https://labs.waterdata.usgs.gov/visualizations/gages-through-the-ages>)
- 963 Velpuri, N. M., and G. B. Senay. "Analysis of long-term trends (1950–2009) in precipitation, runoff and
 964 runoff coefficient in major urban watersheds in the United States." *Environmental Research*
 965 *Letters* 8, no. 2 (2013): 024020. <https://doi.org/10.1088/1748-9326/8/2/024020>
- 966 Vermote, Eric; NOAA CDR Program. (2019): NOAA Climate Data Record (CDR) of AVHRR Leaf Area
 967 Index (LAI) and Fraction of Absorbed Photosynthetically Active Radiation (FAPAR), Version
 968 5. [LAI]. NOAA National Centers for Environmental Information.
 969 <https://doi.org/10.7289/V5TT4P69>.
- 970 Wang, J., Walter, B. A., Yao, F., Song, C., Ding, M., Maroof, A. S., Zhu, J., Fan, C., McAlister, J. M., &
 971 Sikder, S. (2022). GeoDAR: georeferenced global dams and reservoirs dataset for bridging
 972 attributes and geolocations. *Earth System Science Data*, 14(4), 1869-1899.
 973 <https://doi.org/10.5194/essd-14-1869-2022>
- 974 Wilby, R. L., & Dessai, S. (2010). Robust adaptation to climate change. *Weather*, 65(7), 180–185.
 975 <https://doi.org/10.1002/wea.543>
- 976 Xiong, J., Yin, J., Guo, S., He, S., & Chen, J. (2022). Annual runoff coefficient variation in a changing
 977 environment: A global perspective. *Environmental Research Letters*, 17(6), 064006.
 978 <https://doi.org/10.1088/1748-9326/ac62ad>
- 979 Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., & Pavelsky, T. M. (2019). MERIT Hydro:
 980 a high-resolution global hydrography map based on latest topography dataset. *Water Resources*
 981 *Research*, 55(6), 5053-5073. <https://doi.org/10.1029/2019WR024873>
- 982 Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C.,
 983 Kanae, S., & Bates, P. D. (2017). A high-accuracy map of global terrain elevations. *Geophysical*
 984 *Research Letters*, 44(11), 5844-5853. <https://doi.org/10.1002/2017GL072874>
- 985 Yang, L., Yang, Y., Villarini, G., Li, X., Hu, H., Wang, L., Blöschl, G., & Tian, F. (2021). Climate more
 986 important for Chinese flood changes than reservoirs and land use. *Geophysical Research Letters*,
 987 48(11), e2021GL093061. <https://doi.org/10.1029/2021GL093061>
- 988 Yin, Z., Lin, P., Riggs, R., Allen, G. H., Lei, X., Zheng, Z., & Cai, S. (2023). A Synthesis of Global
 989 Streamflow characteristics, Hydrometeorology, and catchment Attributes (GSHA) for Large
 990 Sample River-Centric Studies (1.0) [Data set]. Zenodo.
 991 <https://doi.org/10.5281/zenodo.8090704>
- 992 Ziyun Yin, Peirong Lin, Ryan Riggs, George H. Allen, Xiangyong Lei, Ziyun Zheng, & Siyu Cai. (2023).
 993 A Synthesis of Global Streamflow characteristics, Hydrometeorology, and catchment Attributes
 994 (GSHA) for Large Sample River-Centric Studies V1.1 (1.1) [Data set]. Zenodo.
 995 <https://doi.org/10.5281/zenodo.10127757>

- 996 Zaitchik, B. F., Rodell, M., & Reichle, R. H. (2008). Assimilation of GRACE terrestrial water storage
997 data into a land surface model: Results for the Mississippi River basin. *Journal of*
998 *Hydrometeorology*, 9(3), 535-548. <https://doi.org/10.1175/2007jhm951.1>
- 999 Zhang, J., Lin, P., Gao, S., & Fang, Z. (2020). Understanding the re-infiltration process to simulating
1000 streamflow in North Central Texas using the WRF-hydro modeling system. *Journal of*
1001 *Hydrology*, 587, 124902. <https://doi.org/10.1016/j.jhydrol.2020.124902>
- 1002 Zhang, J., Wang, T., & Ge, J. (2015). Assessing vegetation cover dynamics induced by policy-driven
1003 ecological restoration and implication to soil erosion in southern China. *PLoS One*, 10(6),
1004 e0131352. <https://doi.org/10.1371/journal.pone.0131352>
- 1005 Zhang, S., Zhou, L., Zhang, L., Yang, Y., Wei, Z., Zhou, S., Yang, D., Yang, X., Wu, X., & Zhang, Y.
1006 (2022). Reconciling disagreement on global river flood changes in a warming climate. *Nature*
1007 *Climate Change*, 1-8. <https://doi.org/10.1038/s41558-022-01539-7>
- 1008 Zhang, Y., & Liang, S. (2014). Changes in forest biomass and linkage to climate and forest disturbances
1009 over Northeastern China. *Global change biology*, 20(8), 2596-2606.
1010 <https://doi.org/10.1111/gcb.12588>
- 1011 Zhang, Y., Zheng, H., Zhang, X., Leung, L. R., Liu, C., Zheng, C., Guo, Y., Chiew, F. H., Post, D., &
1012 Kong, D. (2023). Future global streamflow declines are probably more severe than previously
1013 estimated. *Nature Water*, 1-11. <https://doi.org/10.1038/s44221-023-00030-7>
- 1014
- 1015