CCAM: China Catchment Attributes and Meteorology dataset

Zhen Hao[2, *], Jin Jin[1,2, *], Runliang Xia[2], Shimin Tian[2], Wushuang Yang[2], Qixing Liu[2], Min Zhu[2], Tao Ma[2], Chengran Jing[2]

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an, China, 710072
[2] Yellow River Institute of Hydraulic Research, Zhengzhou, China, 450003
*These authors contributed equally to this work.

Correspondence to: Jin Jin (jinjinhao@21cn.com)

Abstract. The lack of a complied large-scale catchment characteristics dataset is a key obstacle limiting the development of large sample hydrology research in China. We introduce the first large-scale catchment attributes dataset in China. We compiled diverse data sources, including soil, land cover, climate, topography, and geology, to develop the dataset. The dataset also includes catchment scale 31-year meteorological time series from 1990 to 2020 for each basin. Potential evapotranspiration time series based on Penman's equation is derived for each basin. The 4911 catchments included in the dataset covers the entire China. We introduced several new indicators describing the catchment geography and the underlying surface compared with previously proposed datasets. The resulting dataset has a total of 125 catchment attributes. The proposed dataset also includes a separate HydroMLYR dataset containing standardized weekly averaged streamflow for 102 basins in the Yellow River Basin. The standardized streamflow data should be able to support machine learning hydrology research in the Yellow River Basin. The proposed dataset is freely available at http://doi.org/10.5281/zenodo.5137288. In addition, the accompanying code for generating the dataset is freely available at https://github.com/haozhen315/CCAM-China-Catchment-Attributes-and-Meteorology-dataset, supporting the generation of catchment characteristics for any custom basin boundaries. Complied data for the 4911 basins covering the entire China and the open-sourced code should be able to support the study of any arbitrary basins instead of being limited to only a few basins.

1 Introduction

Rainfall, interception, evaporation and evapotranspiration, groundwater flow, subsurface flow and surface runoff are the main components of the terrestrial hydrological cycle. These processes are affected by the nature of the catchment, such as the ability of the soil to hold water. Catchment attributes influence the water movement and storage of the catchment such that hydrologic behaviours can vary across catchments (van Werkhoven, Wagener et al. 2008). Studying a large set of terrestrial catchments often provides insights that cannot be obtained when looking at a single or few (Coron, Andreassian et al. 2012, Kollat, Reed et al. 2012, Newman, Clark et al. 2015, Lane, Coxon et al. 2019). For example, a calibrated model may not be applicable in a watershed with vastly different properties. However, by examining a large sample of catchments, it is possible for a data-driven model to learn the similarities and differences of hydrological behaviours across catchments (Kratzert, Klotz et al.

2019). Prediction in ungauged basins is a challenging problem present in hydrology. The central challenge is how to extrapolate hydrologic information from gauged basins to ungauged ones, solving the problem relies on understanding the similarities and differences between different catchments. Regionally, and temporally imbalanced observations bring a difficulty to the problem. For a model to successfully simulate the ungauged areas, it must adapt itself to the different hydrologic behaviours present in different catchments. Kratzert, Klotz et al. (2019) shows encoding catchment characteristics (e.g., soil characteristics, land cover, topography) into a data-driven model can guide the model to behave differently responding to the meteorological time series input based on different sets of catchment attributes.

Large sample hydrological datasets are the foundation and key of many hydrological studies (Silberstein 2006, Shen, Laloy et al. 2018, Nevo, Anisimov et al. 2019). The term big hydrologic data refers to all data influencing the water cycle, such as the meteorological variables, infiltration characteristics of the study area, land use or land cover types, physical and geological features of the study catchment, etc. Many studies are based on large-scale hydrologic data (Coron, Andreassian et al. 2012, Singh, van Werkhoven et al. 2014, Berghuijs, Aalbers et al. 2017, Gudmundsson, Leonard et al. 2019, Tyralis, Papacharalampous et al. 2019). For hydrological research, basin orientated datasets are of great significance. For example, comparative hydrology (de Araújo and González Piedra 2009, Singh, Archfield et al. 2014) focus on understanding how hydrological processes interact with the ecosystem, in particular, how hydrologic behaviours change under changes in the surface and sub-surface of the earth to determine to what extent hydrological predictions can be transferred from one area to another. Large-sample catchment attributes datasets provide opportunities for research studying interrelationships among catchment attributes. Seybold, Rothman et al. (2017) studied the correlations between river junction angle with geometric factors, downstream concavity, and aridity. Oudin, Andréassian et al. (2008) investigates the link between land cover and mean annual streamflow based on 1508 basins representing a large hydroclimatic variety. Voepel, Ruddell et al. (2011) examines how the interaction of climate and topography influences vegetation response.

World-wide data sharing has become a trend (Wickel, Lehner et al. 2007, Ceola, Arheimer et al. 2015, Blume, van Meerveld et al. 2018, Wang, Chen et al. 2020), and the amounts of hydrologic data available are ever-increasing. However, these data typically came from different providers and are compiled in various formats. ASTGTM (Abrams, Crippen et al. 2020) provides a global digital elevation model; GliM (Hartmann and Moosdorf 2012) includes rock types data globally; MODIS provides data products (Knyazikhin 1999, Didan 2015, Myneni, Knyazikhin et al. 2015, Running, Mu et al. 2017, Sulla-Menashe and Friedl 2018) describing features of the land and the atmosphere derived from remote sensing observations; Yamazaki, Ikeshima et al. (2019) provides a global flow direction map at three arc-second resolution; HydroBASINS (Lehner 2014) provides basin boundaries at different scales globally; and GDBD (Masutomi, Inui et al. 2009) provides basin boundaries with geographic attributes; GLHYMPS (Gleeson, Moosdorf et al. 2014) provides a global map of subsurface permeability and porosity; SoilGrids250m (Hengl, Mendes de Jesus et al. 2017) dataset provides global numeric soil properties. Local government agencies often hold meteorological data such as precipitation and evaporation, and the amount of this data is also growing.

However, the data mentioned above are rarely spatially aggregated to the catchment scale, making it difficult for researchers to use these data. Properly pre-processed and formatted datasets are of great importance for hydrology research. Searching for appropriate data sources, pre-processing, and formatting often consumes a lot of time. In some cases, individual research groups either do not know where to obtain the appropriate data or cannot properly process the data to receive the desired format. In summary, although data sharing is being advocated in the community, it is usually difficult for the public to obtain the required data, either because there are not enough observations or because of the difficulties in the data processing.

Recently, there are efforts (Addor, Newman et al. 2017, Alvarez-Garreton, Mendoza et al. 2018, Chagas, Chaffe et al. 2020, Coxon, Addor et al. 2020) to compile different types of data sources forming large scale hydrological datasets. These four collected datasets cover the continental United States, Chile, Brazil, and Great Britain. Addor, Do et al. (2020) reviewed these datasets and discussed the guidelines for producing large-sample hydrological datasets and the limitations of the currently proposed datasets. The static properties of 671 river basins in the United States are calculated by CAMELS (Addor, Newman et al. 2017), which is an extension of a previously proposed hydrometeorological data set (Newman, Clark et al. 2015). Unfortunately, it is impossible to publish streamflow data in China for the time being. The CAMELS dataset has been used to support a lot of research. For example, Knoben, Freer et al. (2019) compared metrics used in hydrology based on simulations on many basins. Tyralis, Papacharalampous et al. (2019) studied the relationship between the shape parameter and basin attributes based on the sizeable basin-oriented dataset.

There is currently no compilation of China-specific catchment attributes datasets. An alternative, the HydroATLAS (Linke, Lehner et al. 2019) dataset, which is on a global scale, is basically performing zonal statistics on the source data. HydroATLAS lacks many indicators which need derivations based on the source data, such as rainfall seasonality, the fraction of precipitation falling as snow, basin shape factors and root depth distributions. What’s worse, the meteorological data is only up to 2000, which is outdated.

In summary, a lack of a complied catchment attributes dataset is a key obstacle limiting the development of large sample hydrology research in China. Inspired by (Addor, Newman et al. 2017), we complied multiple data sources, including basin topography, climate indices, land cover characteristics, soil characteristics and geological characteristics. Different from (Addor, Newman et al. 2017), the catchments included in the dataset covers the entire study area, instead of being limited to a few.

The proposed dataset is the first dataset providing catchments meteorological time series and catchments attributes of China. We compiled and named the dataset following most standards of the previously proposed datasets. The dataset consists of all derived basin boundaries from the Digital Elevation Model (DEM) which came from the Global Drainage Basin Dataset (Masutomi, Inui et al. 2009). The Global Drainage Basin Dataset (GDBD) is derived at high-resolution (100m-1km) and has a good geographic agreement with existing global drainage basin data in China. In addition, previously proposed datasets (Addor, Newman et al. 2017, Alvarez-Garreton, Mendoza et al. 2018, Chagas, Chaffe et al. 2020, Coxon, Addor et al. 2020) report only the most frequent catchment land cover and lithology types. Instead, CCAM calculates the proportions of all land cover and lithology types.

In addition to the basin-wise attributes provided in CCAM, we propose HydroMLYR, a hydrology dataset for machine learning research in the Yellow River Basin providing weekly averaged standardized streamflow data for 102 basins in the Yellow

River Basin (YRB). HydroMLYR is proposed to support machine learning hydrology research at YRB. Traditional hydrological models have some long standing challenges, such as the inability to capture hydrological processes' mechanism complexity (Kollat, Reed et al. 2012), which is due to the structural limitations of the conceptual models. Data-driven strategies represented by machine learning are proposed to overcome some existing obstacles and they open a new way for researchers to acquire knowledge transforming the research pattern from hypothesis-driven to data-driven. Feng, Fang et al. (2020) proposed a flexible data integration fusing various types of observations to improve rainfall-runoff modelling. The research shows that combining different resources of data benefits predictions in regions with high autocorrelation in streamflow. Wongso, Nateghi et al. (2020) developed a model predicting the state-level, per capita water uses in the United States, taking various geographic, climatic, and socioeconomic variables as input. The research also identified key factors associated with high water usage. Mei, Maggioni et al. (2020) proposed a statistical framework for spatial downscaling to obtain hyper-resolution precipitation data. The results show improvements compared with the original product. Brodeur, Herman et al. (2020) applied machine learning techniques, namely bootstrap aggregation and cross-validation, to reduce overfitting in reservoir control policy search. Ni and Benson (2020) proposed an unsupervised machine learning method to differentiate flow regimes and identify capillary heterogeneity trapping, showing the promise of machine learning methods for analysing large datasets from coreflooding experiments. Legasa and Gutiérrez (2020) propose to apply Bayesian Network for multisite precipitation occurrence generation, and the proposed methodology shows improvements for existing methods. The proposed data set can be used to develop or verify machine learning models in the YRB.

The paper is organized as follows: Section 2 describes the study area. Section 3-7 describes the five classes of the computed catchment attributes. Section 8 describes the proposed catchment scale meteorological time series. Section 9 introduce the HydroMLYR dataset. Section 10 describes the code and data availability. Section 11 is the concluding remark.

2 Study area

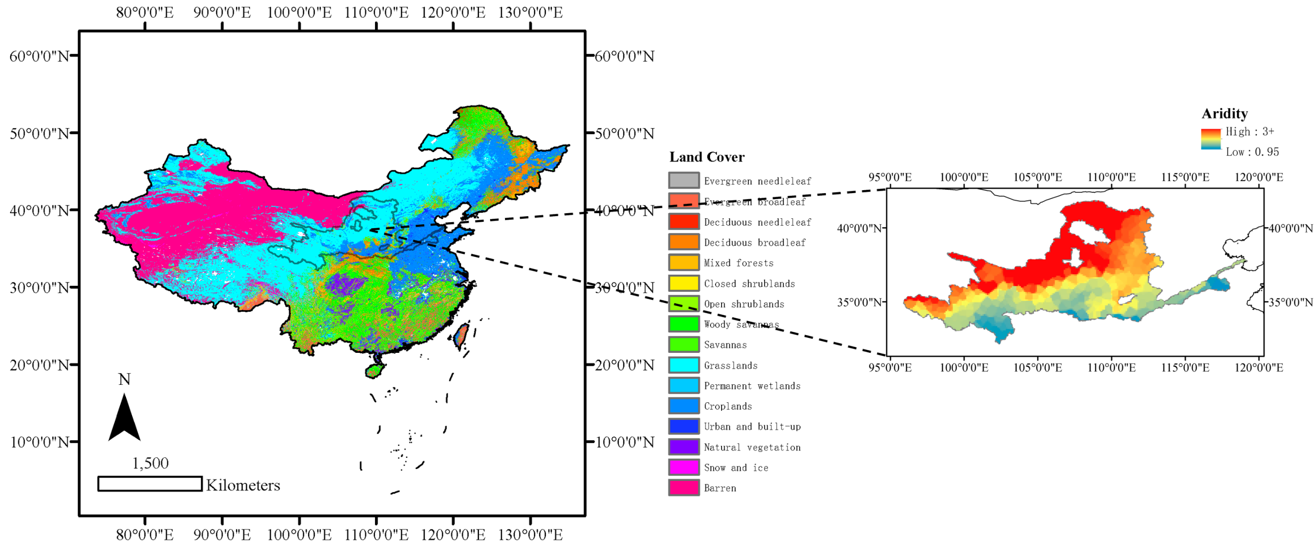


Figure 1: Left: Study area of CCAM and the distribution of land cover types. The studied basins cover the whole of China. Right: Study area of HydroMLYR and the distribution of aridity (PET/P) index. YRB is a generally arid area. The data set provided can be used as a good sample for studying hydrology in arid regions.

The study area corresponds to the whole of China (Fig. 1), with diverse climate and terrain characteristics, spanning from 18.2° N to 52.3° N and 76.0° E to 134.3° E. Mountains, plateaus, and hills account for about two-thirds of areas of China, and the remaining are basins and plains. China’s topography is like a three-level ladder, high in the west and low in the east. The Qinghai-Tibet Plateau, the highest plateau globally, located in the west of China, with a mean elevation of over 4000 meters, is the first step of China’s topography. The Xinjiang region, the Loess Plateau, the Sichuan Basin, and the Yunnan-Guizhou Plateau to the north and east are the second step of China’s topography. The mean sea level here is between 1000 to 2000 meters. Plains and hills dominate the east of the Daxinganling-Taihang Mountain to the coastline, the third step of China. The elevation of this step descends to 500-1,000 meters. To better characterize the studied catchments, we have derived various attributes. Table 1 compares the number of derived attributes between several proposed datasets.

Table 1: Number of computed attributes in CAMELS, CAMELS-BR and CCAM.

Attribute class	CAMELS(A17)	CAMELS-BR	CCAM
Location and topography	9	11	12
Geology	7	7	18
Soil	11	6	54
Land cover	8	11	22
Climatic indices	11	13	17

Human intervention indices	not computed	4	2
Total	46	52	125

In China, precipitation and temperature vary significantly in different places, forming a diverse climate environment. According to the Köppen Climate Classification System, from northwest to southeast, China's climate gradually evolves from Cold desert (BW_k) climate, Tundra (ET) climate, Warm and temperate continental (D_{fa} and D_{wb}) climate to Humid subtropical (C_{wa}) climate and Warm oceanic (C_{fa}) climate. From the perspective of temperature zones, there are tropical, subtropical, warm temperate, medium temperate, cold temperate and Qinghai-Tibet Plateau regions, and there are humid regions, semi-humid regions, semiarid regions, and arid regions from the perspective of wet and dry zones. Moreover, the same temperature zone can contain different dry and wet zones. Therefore, there will be differences in heat and wetness in the same climate type. The complexity of the terrain makes the climate even more complex and diverse. Besides, China has a wide range of regions affected by the alternating winter and summer monsoons. Compared with other parts of the world at the same latitude, these areas have low winter temperatures, high summer temperatures, significant annual temperature differences, and concentrated precipitation in summer. The cold and dry winter monsoon occurs in Asia's interior, far away from the ocean. Under its influence, winter rainfall in most parts of China is low, accompanied by low temperature. The summer monsoon is warm and humid, coming from the Pacific Ocean and the Indian Ocean. Under its influence, precipitation generally increases. Table 2 compares the provided forcing variables in CAMELS, CAMELS-BR and CCAM.

Table 2: Summary of forcing variables provided in CAMELS, CAMELS-BR and CCAM.

Forcing data class	CAMELS	CAMELS-BR	CCAM
Temperature	available	available	available
Precipitation	available	available	available
Solar radiation	available	**not available**	available
Day length	available	**not available**	**not available**
Sunshine hours	**not available**	**not available**	available
Humidity	available	**not available**	available
Snow water equivalent	available	**not available**	**not available**
Wind velocity	**not available**	**not available**	available
Ground surface pressure	available	**not available**	available
Observed evaporation	**not available**	available	available
Potential evapotranspiration	**not available**	available	available

Table 3: Summary table of catchment attributes available in the proposed dataset.

Attribute class	Attribute name	Description	Unit	Data source
Climate indices (computed for 1 Oct 1990 to 30 Sep 2018)	pet_mean	mean daily pet (Penman–Monteith equation)	mm d^{-1}	(Subramanya 2013)
	evp_mean	mean daily evaporation (observations)	mm d^{-1}	SURF_CLI_CHN_MUL_DAY3F[1]
	gst_mean	mean daily ground surface temperature	°C	
	pre_mean	mean daily precipitation	mm d^{-1}	
	prs_mean	mean daily ground surface pressure	hPa	
	rhu_mean	mean daily relative humidity	-	
	ssd_mean	mean daily sunshine duration	h	
	tem_mean	mean daily temperature	°C	
	win_mean	mean daily wind speed	m s^{-1}	
	p_seasonality	seasonality and timing of precipitation (estimated using sine curves to represent the annual temperature and precipitation cycles, positive [negative] values indicate that precipitation peaks in summer [winter], values close to 0 indicate uniform precipitation throughout the year)	-	
	high_prec_freq	frequency of high-precipitation days (⩾ 5 times mean daily precipitation)	d yr^{-1}	
	high_prec_dur	average duration of high-precipitation events (number of	d	

[1] http://data.cma.cn/data/cdcdetail/dataCode/SURF_CLI_CHN_MUL_DAY.html

		consecutive days ⩾ 5 times mean daily precipitation)		
	high_prec_timing	season during which most high-precipitation days (⩾ 5 times mean daily precipitation) occur	season	
	low_prec_freq	frequency of dry days (< 1mm d^{-1})	d yr^{-1}	
	low_prec_dur	average duration of dry periods (number of consecutive days < 1 mm d^{-1})	d	
	low_prec_timing	season during which most dry days (< 1 mm d^{-1}) occur	season	
	frac_snow_daily	fraction of precipitation falling as snow (for days colder than 0 °C)	-	
	p_seasonality	seasonality and timing of precipitation, positive [negative] values indicate that precipitation peaks in summer [winter], values close to 0 indicate uniform precipitation throughout the year	-	
Geological characteristics	geol_porosity	subsurface porosity	-	(Gleeson, Moosdorf et al. 2014)
	geol_permeability	subsurface permeability (log-10)	m^2	
	ig	fraction of the catchment area associated with ice and glaciers	-	(Hartmann and Moosdorf 2012)
	pa	fraction of the catchment area associated with acid plutonic rocks	-	
	sc	fraction of the catchment area associated with carbonate sedimentary rocks	-	
	su	fraction of the catchment area associated with unconsolidated sediments	-	

sm	fraction of the catchment area associated with mixed sedimentary rocks	-
vi	fraction of the catchment area associated with intermediate volcanic rocks	-
mt	fraction of the catchment area associated with metamorphic	-
ss	fraction of the catchment area associated with siliciclastic sedimentary rocks	-
pi	fraction of the catchment area associated with intermediate plutonic rocks	-
va	fraction of the catchment area associated with acid volcanic rocks	-
wb	fraction of the catchment area associated with water bodies	-
pb	fraction of the catchment area associated with basic plutonic rocks	-
vb	fraction of the catchment area associated with basic volcanic rocks	-
nd	fraction of the catchment area associated with no data	-
py	fraction of the catchment area associated with pyroclastic	-
ev	fraction of the catchment area associated with evaporites	-

Land cover characteristics	lai_max	maximum monthly mean of the leaf area index (based on 12 monthly means)	-	(Myneni, Knyazikhin et al. 2015)
	lai_diff	difference between the maximum and minimum monthly mean of the leaf area index (based on 12 monthly means)	-	
	ndvi_mean	mean normalized difference vegetation index (NDVI)	-	(Didan 2015)
	root_depth_50	root depth (percentiles=50% extracted from a root depth distribution based on IGBP land cover)	m	Eq. 2 and Table 2 in (Zeng 2001)
	root_depth_99	root depth (percentiles=99% extracted from a root depth distribution based on IGBP land cover)	m	
	evergreen needleleaf tree	catchment area fraction covered by evergreen needleleaf tree	-	(Sulla-Menashe and Friedl 2018)
	evergreen broadleaf tree	catchment area fraction covered by evergreen broadleaf tree	-	
	deciduous needleleaf tree	catchment area fraction covered by deciduous needleleaf forests	-	
	deciduous broadleaf tree	catchment area fraction covered by deciduous broadleaf tree	-	
	mixed forest	catchment area fraction covered by mixed forest	-	
	closed shrubland	catchment area fraction covered by closed shrubland	-	
	open shrubland	catchment area fraction covered by open shrubland	-	
	woody savanna	catchment area fraction covered by woody savanna	-	

	savanna	catchment area fraction covered by savanna	-	
	grassland	catchment area fraction covered by grassland	-	
	permanent wetland	catchment area fraction covered by permanent wetland	-	
	cropland	catchment area fraction covered by cropland	-	
	urban and built-up land	catchment area fraction covered by urban and built-up land	-	
	cropland/natural vegetation	catchment area fraction covered by cropland/natural vegetation	-	
	snow and ice	catchment area fraction covered by snow and ice	-	
	barren	catchment area fraction covered by barren	-	
	water bodies	catchment area fraction covered by water bodies	-	
Topography, location and Human intervention	basin_id	drainage basin identifiers	-	(Masutomi, Inui et al. 2009)
	pop	population	people	
	pop_dnsty	population density	people km^{-2}	
	lat	mean latitude	°N	
	lon	mean longitude	°E	
	elev	mean elevation	M	
	area	catchment area	km^2	
	slope	mean slope	m km^{-1}	(Horn 1981)
	length	The length of the mainstream measured from the basin outlet to the remotest point on the basin boundary. The mainstream is identified by starting from the	Km	(Subramanya 2013)

		basin outlet and moving up the catchment.		
	form factor	catchment area / (catchment length)^2	-	
	shape factor	(catchment length)^2 / catchment area	-	
	compactness coefficient	perimeter of the catchment / perimeter of the circle whose area is that of the basin	-	
	circulatory ratio	catchment area / area of circle of catchment perimeter	-	
	elongation ratio	diameter of circle whose area is basin area / catchment length	-	
Soil	pdep	soil profile depth	cm	(Shangguan, Dai et al. 2013)
	clay	percentage of clay content of the soil material	%	
	sand	percentage of sand content of the soil material	%	
	por	porosity	cm^3 cm^{-3}	
	silt	percentage of silt content of the soil material	%	
	grav	rock fragment content	%	
	som	soil organic carbon content	%	
	log_k_s4F[2]	log-10 transformation of saturated hydraulic conductivity	cm d^{-1}	(Dai, Xin et al. 2019)
	theta_s[4]	saturated water content	cm^3 cm^{-3}	
	tksatu[4]	thermal conductivity of unfrozen saturated soils	W m^{-1} K^{-1}	
	bldfie[4]	bulk density	kg m^{-3}	

[2] The data source contains multi-layer soil data, soil characteristics for all layers are determined.

	cecsol[4]	cation-exchange capacity	cmol+ kg^{-1}	(Hengl, Mendes de Jesus et al. 2017)
	orcdrc[4]	organic carbon content	g kg^{-1}	
	phihox[4]	pH in H2O	10^{-1}	
	bdticm	depth to bedrock	cm	

3 Climatic indices

Raw meteorological data is provided by the China Meteorological Data Network, released as the SURF_CLI_CHN_MUL_DAY (V3.0) dataset[3], which provides the longest period (1951-2020) of meteorological time series in China. The SURF_CLI_CHN_MUL_DAY product includes site observations of pressure, temperature, relative humidity, precipitation, evaporation, wind speed, sunshine duration, and ground surface temperature (Table 4). The Inverse distance weighting method is used for interpolating the site observations. To ensure data quality, we use the latter 31-year record (from 1990 to 2020) to construct the dataset since sites' distribution was sparse in the early days (Fig. 2). We computed more climatic characteristics compared with other datasets (Table 2). These variables are useful in hydrological modelling; for example, wind speed can affect actual evapotranspiration. To be consistent with the CAMELS (Addor, Newman et al. 2017), we determined all climatic attributes (Woods 2009) provided in the CAMELS dataset. As a result, the proposed dataset provides more meteorological variables and longer time series (1990-2020) than CAMELS and CAMELS-CL. A summary of the derived climate indices is presented in Table 3. The national distributions of the climate indicators are shown in Fig. 3.

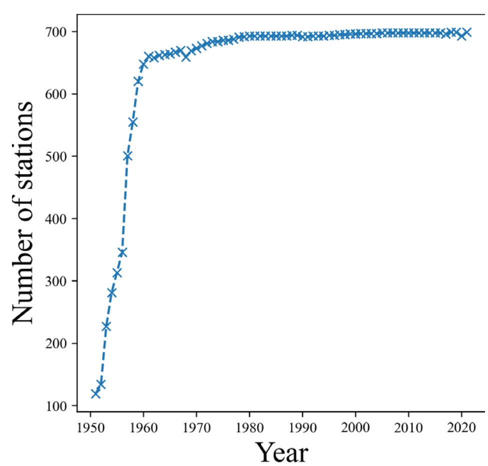
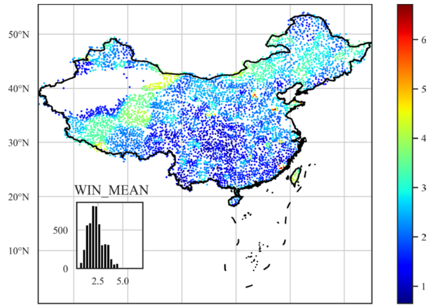
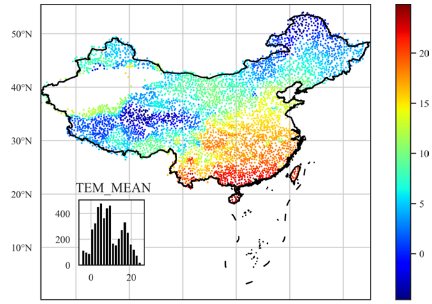


Figure 2: Changes in the number of meteorological stations in China. There were only 119 stations in 1951. This number increased rapidly from 1951 to the early 1960s, and the number of stations remained stable after 2000. To ensure the data quality, we used the latter 31-year records (from 1990 to 2020) to construct the dataset.

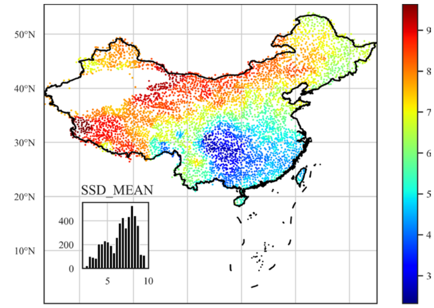
[3] SURF_CLI_CHN_MUL_DAY is freely available for global researchers.



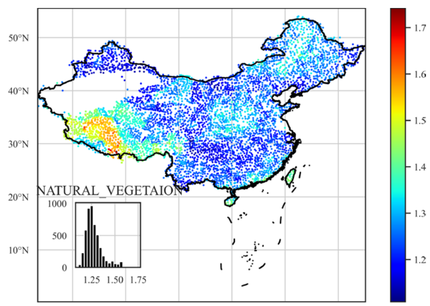
(a)



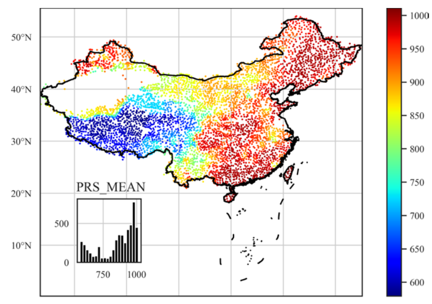
(b)



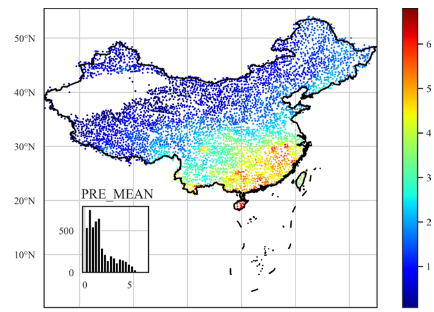
I



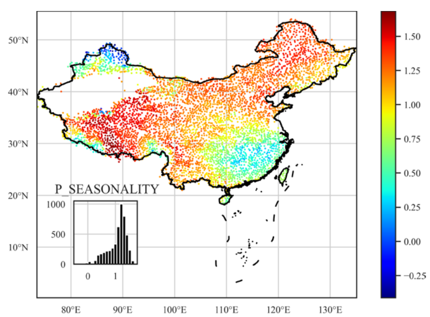
(d)



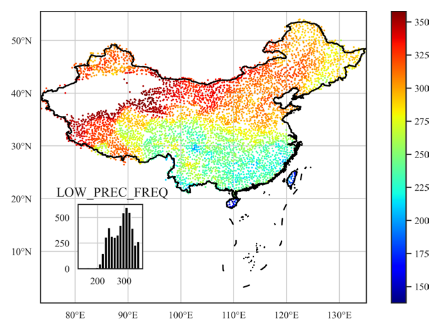
I



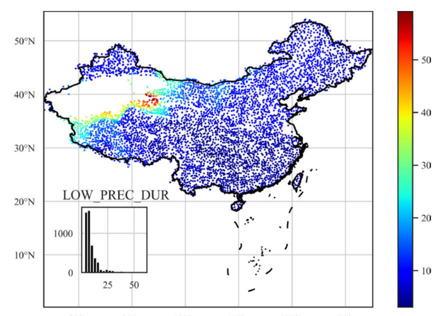
(f)



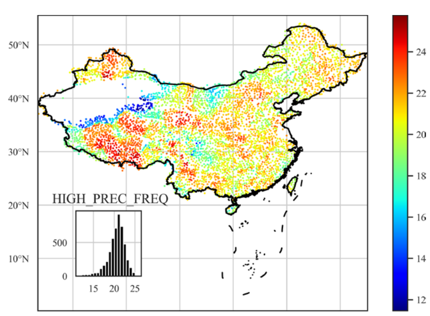
(g)



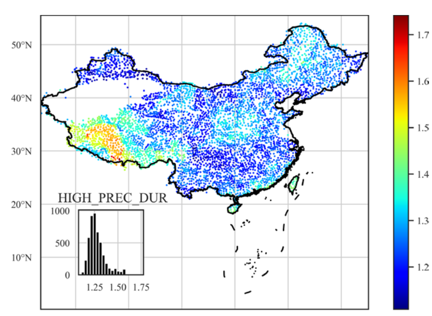
(h)



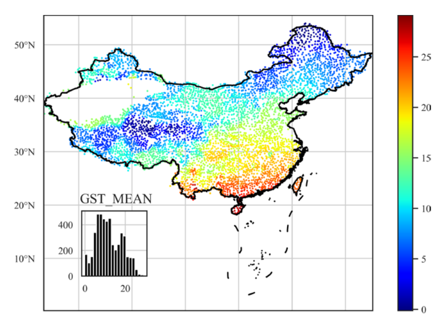
(i)



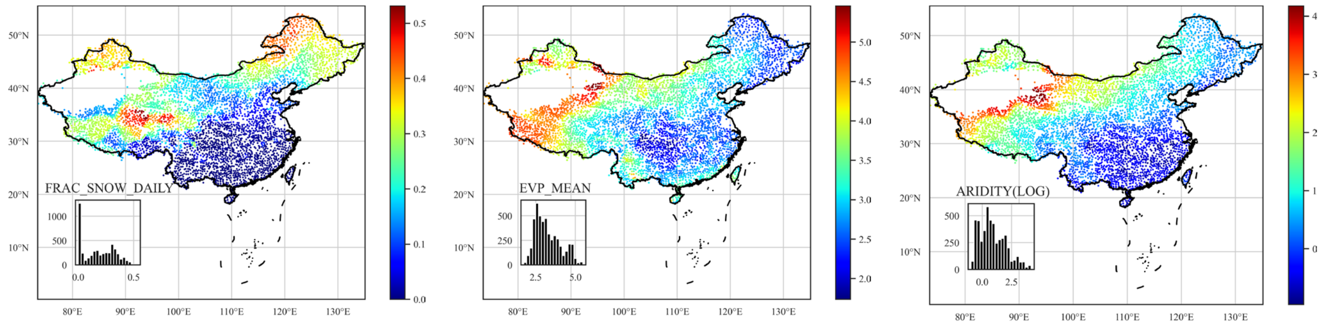
(j)



(k)



(l)



(m) (n) (o)

Figure 3: Distributions of climatic indices over China. All basins are plotted in the same size. When extreme values of a variable affect visualization (cause most areas to have the same colour), the log values are used for visualization.

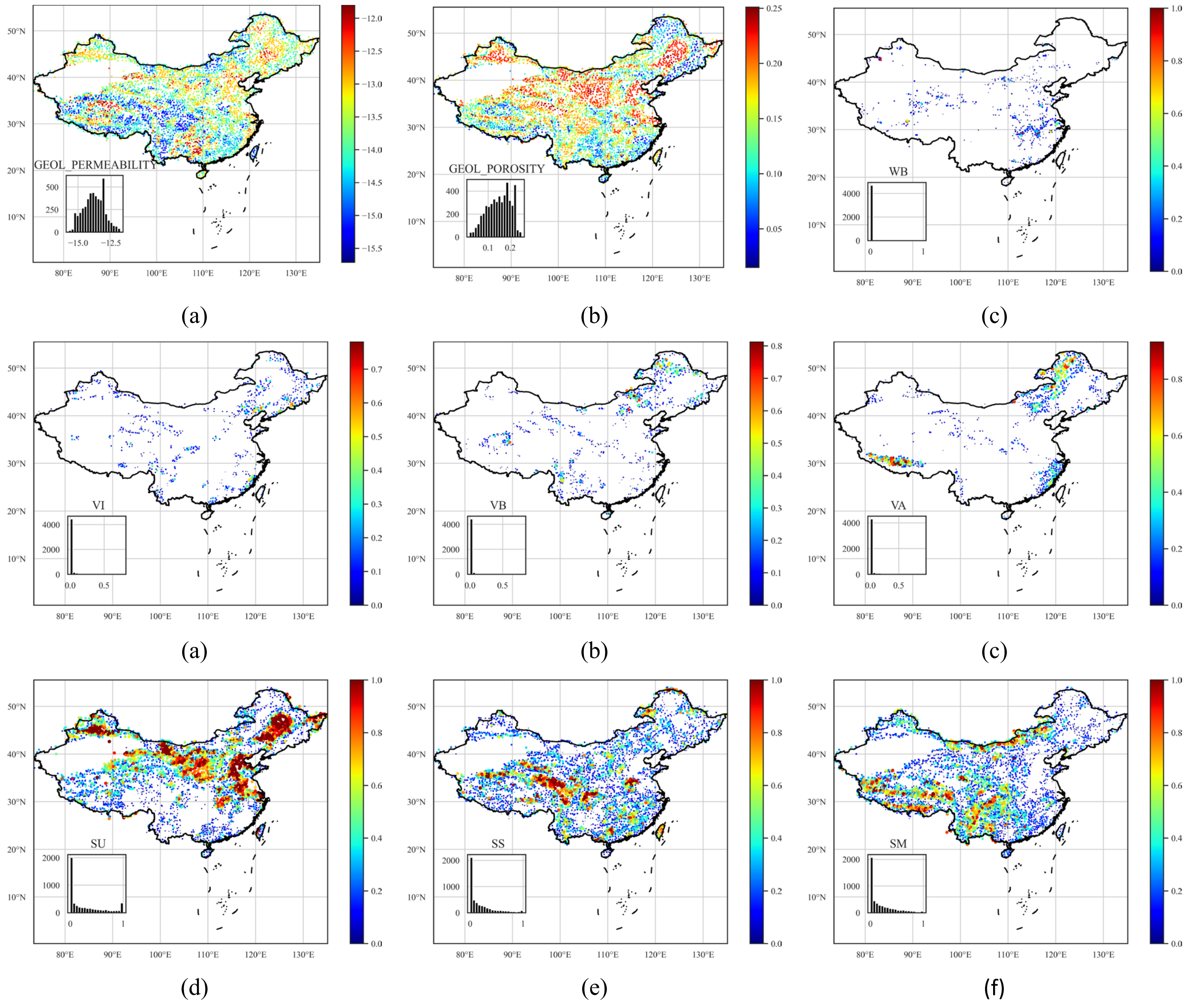
The instruments for measuring potential evaporation were updated from 2000 to 2005. Early observations can be multiplied by a correction coefficient to approximate the new tools. However, the coefficient varies across stations making the approach infeasible. To complement this, we calculated potential evapotranspiration (PET) based on a modified Penman's Equation (Appendix A) and other observed meteorological variables, providing a series of consistent potential evaporation estimations for reference.

The average daily precipitation in China is highest in the southeast and lowest in the northwest. It is also higher in the coastal areas than in the interior land. Ground surface pressure is positively correlated with elevation, the highest in the Qinghai-Tibet Plateau and the lowest in the Southeast Plain. The average relative humidity is generally positively correlated with precipitation; they are also higher in some forested areas, such as the Taihang Mountains and Daxingan Mountains. The Qinghai-Tibet Plateau has the lowest average temperature, and the southern coastal area has the highest. A distinctive feature of the distribution of wind speed is the high wind speed in mountainous areas. The highest wind speed occurs in the southeast coastal area (> 6 meters per second).

4 Geology

To describe the lithological characteristics of each catchment, we used the same two global datasets as CAMELS, Global Lithological Map (GliM) (Hartmann and Moosdorf 2012) and Global Hydrogeology MaPS (GLHYMPS) (Gleeson, Moosdorf et al. 2014). Figure 4 presents the distributions of the geological types.

GLiM provides a high resolution global lithological map assembled from existing regional geological maps; it has been widely used for constructing datasets (e.g. SoilGrids250m (Hengl, Mendes de Jesus et al. 2017)). However, the data quality of GLiM can vary in different spatial locations depending on the quality of the original regional geological maps. GLiM consists of three levels, the first level contains 16 lithological classes, and the additional two levels describe more specific lithological characteristics. The GLiM is represented by 1,235,400 polygons; the polygons are converted to raster format for the basin-scale lithological type statistics. For China, the compiled regional data sources (China 1991, Xinjiang 1992, Survey 2001) have slightly lower resolutions than the GLiM target resolution (1:1 000 000). However, for a basin-scale study with a mean basin area of over 2000 km^2, the classification accuracy should satisfy most applications. Different from CAMELS and CAMELS-CL, we determined each lithological class's contribution to the catchment instead of recoding just the first and second most frequent classes.



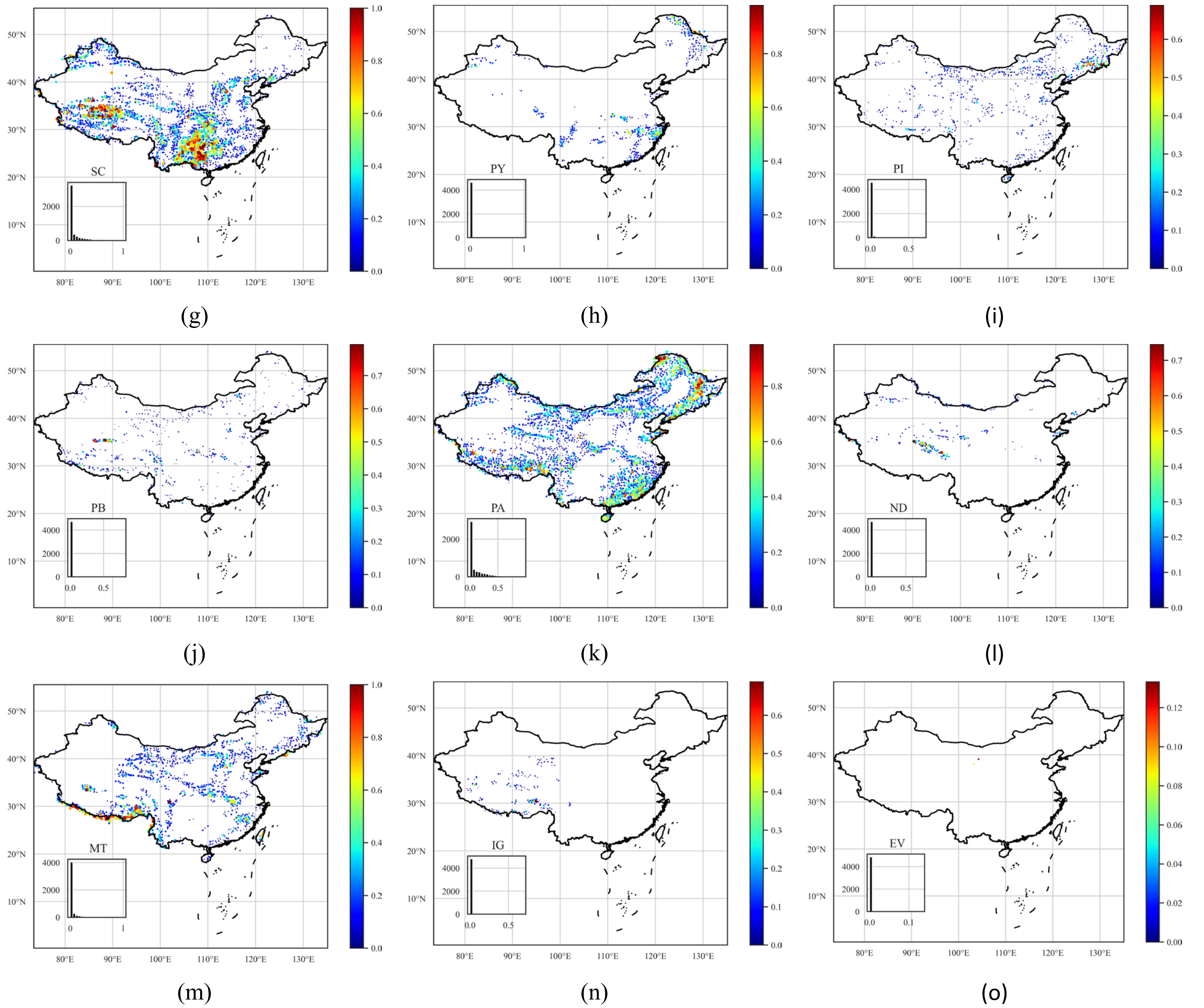


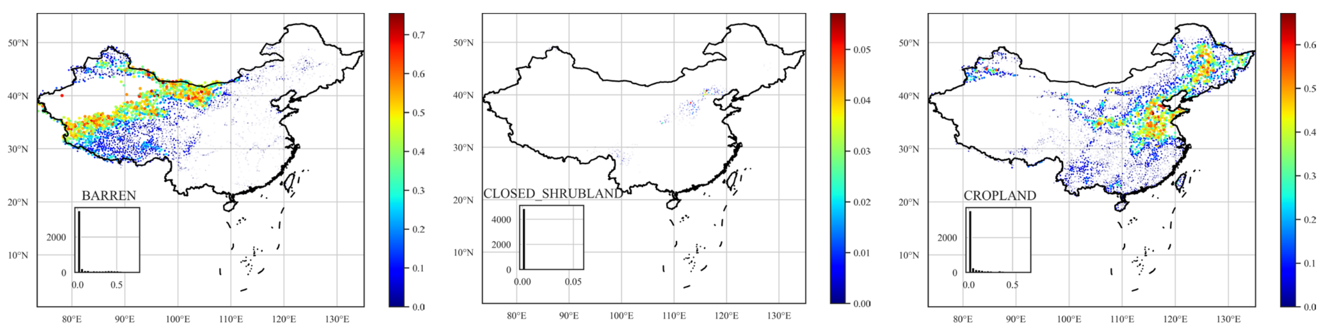
Figure 4: Distributions of geological characteristics over China. For lithologies, the plot size is scaled by the lithology proportion.

GLobal HYdrogeology MaPS (GLHYMPS) provides a global estimation of subsurface permeability and porosity, two critical characteristics for the soils' hydrological classification. Porosity and permeability influence an area's infiltration capacity. Soil with high porosity is likely to contain s amounts of water, and high permeable soil transmits water relatively quickly. Based on the high-resolution map of GLiM, which can differentiate fine and coarse-grained sediments and sedimentary rocks, GLHYMPS determined subsurface permeability depending on the different permeabilities of rock types. For the proposed dataset, we calculated the catchment arithmetic mean for porosity. Followed (Gleeson, Smith et al. 2011), the logarithmic scale geometric mean is used for representing subsurface permeability. The summary of geological characteristics is present in Table 3.

Porosity and permeability have similar distributions as geological classes. These two characteristics are highly dependent on rock properties, unconsolidated sediments, mixed sedimentary rocks, siliciclastic sedimentary rocks, carbonate sedimentary rocks, and acid plutonic rocks are the five most common geological classes in China. Unconsolidated sediment is the most common rock type in China, dominating 31.9% of catchments; it extends from Xinjiang to the inland of the northeast and the coastal area surrounding the Bohai Sea, due to the high proportion of unconsolidated sediments present in the rock, these areas typically have high permeability and medium porosity. Mixed sedimentary rocks are the second most common rock type in China, accounting for 20.3% of catchments, it dominated the southern Qinghai-Tibet Plateau, western Yunnan-Guizhou Plateau, and northern Inner Mongolia. These areas typically have high porosity and low permeability. Siliciclastic sedimentary rocks dominate 17.7% of basins, mainly distributed in the northern part of the Qinghai-Tibet Plateau and the junction of the Qinghai-Tibet Plateau and the Yunnan-Guizhou Plateau; there are also some distributions in the eastern inland. These areas have low subsurface permeability and high subsurface porosity. Amongst all catchments, 9.8% of catchments are dominated by carbonate sedimentary rocks. Carbonate sedimentary rocks are mainly located in eastern Yunnan and northern Qinghai-Tibet Plateau. Acid plutonic rocks are typically distributed in the mountains surrounding the inland northeast, namely the Daxinganling Mountain and the hills in southern Guangdong and southwestern Guangxi. They are also distributed along the Brahmiputra river in the south part of the Qinghai-Tibet Plateau. The distribution of Acid plutonic rocks is relatively scattered; there are many isolated Acid plutonic rocks distributions in different locations of China, accompanied by medium permeability and high porosity.

The types of rocks in China are dominated by unconsolidated sediments and mixed sedimentary rocks. In 33.86% of the catchments, the dominant rock types occupy less than 50% of the catchment areas, and only 16.8% of basins are having a dominant rock type with an area fraction greater than 90%. Amongst 4911 basins, 9.4% of basins have prevalent rock types wholly occupying the area.

5 Landcover



(a) (b) (c)

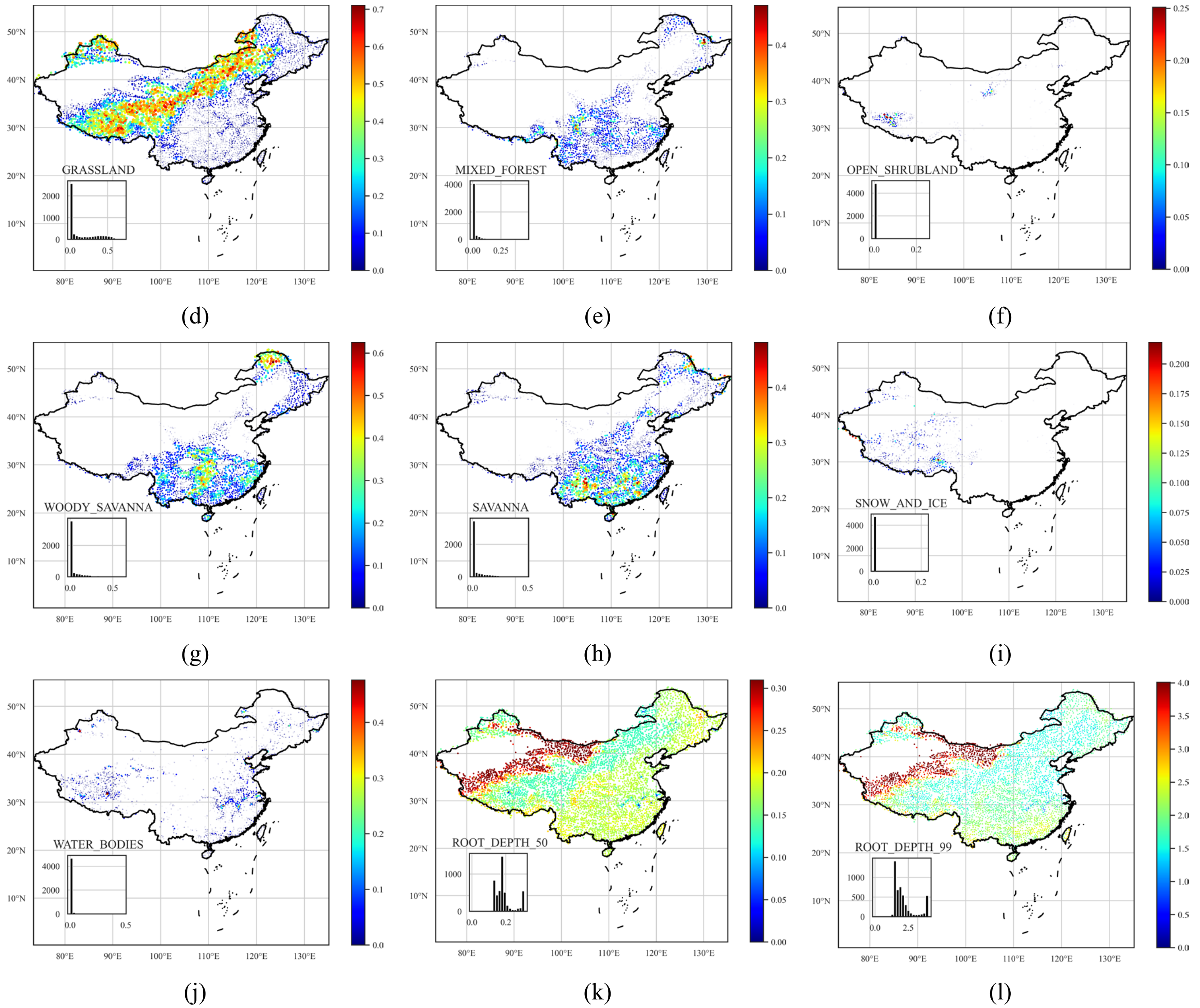


Figure 5: Distributions of land cover characteristics over China. For land cover types, the plot size is scaled by the size of the land cover proportion.

We selected two indicators to characterize vegetation density and growth on the surface: Normalized difference vegetation index (NDVI) and Leaf area index (LAI). NDVI is an indicator with a valid range of -0.2 to 1, assessing whether the area being observed contains live green vegetation or the plants' health. However, NDVI is just a qualitative measurement of the vegetation density; it cannot provide a quantitative estimate of the vegetation density in the area. Moreover, NDVI often provides inaccurate vegetation density measurements, and only long-term measurement and comparison can ensure its accuracy. NDVI alone is not enough to estimate the state of plants in an area. Therefore, we have selected another indicator, LAI, to supplement the deficiencies of NDVI.

LAI is defined as the total needle surface area per unit ground area and half of the entire needle surface area per unit ground surface area. It is a quantifiable value. It is functionally related to many hydrological processes like water interception (van Wijk and Williams 2005). (Buermann, Dong et al. 2001) verifies the validity of LAI used to characterize vegetation growth. The data sources used are The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Vegetation Indices (Didan 2015) for NDVI and Moderate Resolution Imaging Spectroradiometer (MODIS) (Myneni, Knyazikhin et al. 2015) for LAI. Followed (Addor, Newman et al. 2017), we determined maximum monthly LAI as an indicator characterising vegetation interception capacity and the maximum evaporative capacity and the difference between the maximum and minimum monthly LAI representing LAI's temporal variations.

Land cover classification refers to segmenting the ground into different categories based on remote sensing images. The Terra and Aqua combined Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover Type provides different results depending on the classification system used. Annual International Geosphere-Biosphere Programme (IGBP) classification is used for building the dataset, which is derived by the c4.5 decision tree algorithm. The IGBP classification system was formulated by the IGBP Land Cover Working Group in 1995, resulting in 17 categories of land cover types (Belward, Estes et al. 1999). Friedl, Sulla-Menashe et al. (2010) compared the IGBP data of MODIS with other reference datasets and concluded that the MODIS classification of IGBP has an accuracy of 75%. We determined the fraction of each land cover class for each basin based on the Terra and Aqua combined Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover Type (Sulla-Menashe and Friedl 2018), which differentiates our dataset from CAMELS and CAMELS-CL (only calculated the proportion of the dominant types).

Followed (Addor, Newman et al. 2017), we computed the average rooting depth (50% and 90%) for each catchment based on the IGBP classification using a two-parameter method (Zeng 2001). The root depth distribution of vegetation affects the ground's water holding capacity and the topsoil layer's annual evapotranspiration (Desborough 1997). Many models use root depth as an essential parameter to characterize soil moisture absorption capacity. (Zeng 2001) developed a two-parameter asymptotic equation for estimating root depth distribution; the root depth distribution is global, derived based on the IGBP classification avoiding the problem of significantly different root distributions in various research. Figure 5(g) shows root depth distributions of different vegetation types, based on (Zeng 2001). The 90% root depth is usually considered to be "rooting depth", among the 17 categories of IGBP, cropland has the smallest rooting depth, and open shrubland has the largest. The 90% root depth of all vegetation is less than 2 meters. The national distribution of catchments soil characteristics is shown in Fig. 5.

6 Location and topography

The catchments' boundary files are obtained from the global drainage basin dataset (Masutomi, Inui et al. 2009). The GDBD dataset was derived from digital elevation models (DEMs) with a high-resolution (100m-1km), and the errors were corrected by either automatic methods or manually. Additionally, GDBD also provides population and population density estimates for catchments, and these two indicators are also included in our dataset as a measure of human intervention. Global Streamflow Data Centre (Center 2005) discharge gauging stations were used for referencing the derived basins. GDBD has a high average match area rate (AMAR) and good geographic agreement with existing global drainage basin data in China. Based on the high-quality dataset, precise geographic and topographic information can be derived.

The topography attributes of each catchment are determined based on the ASTGTM product retrieved from https://lpdaac.usgs.gov, maintained by the NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC) at the USGS Earth Resources Observation and Science (EROS) Center.

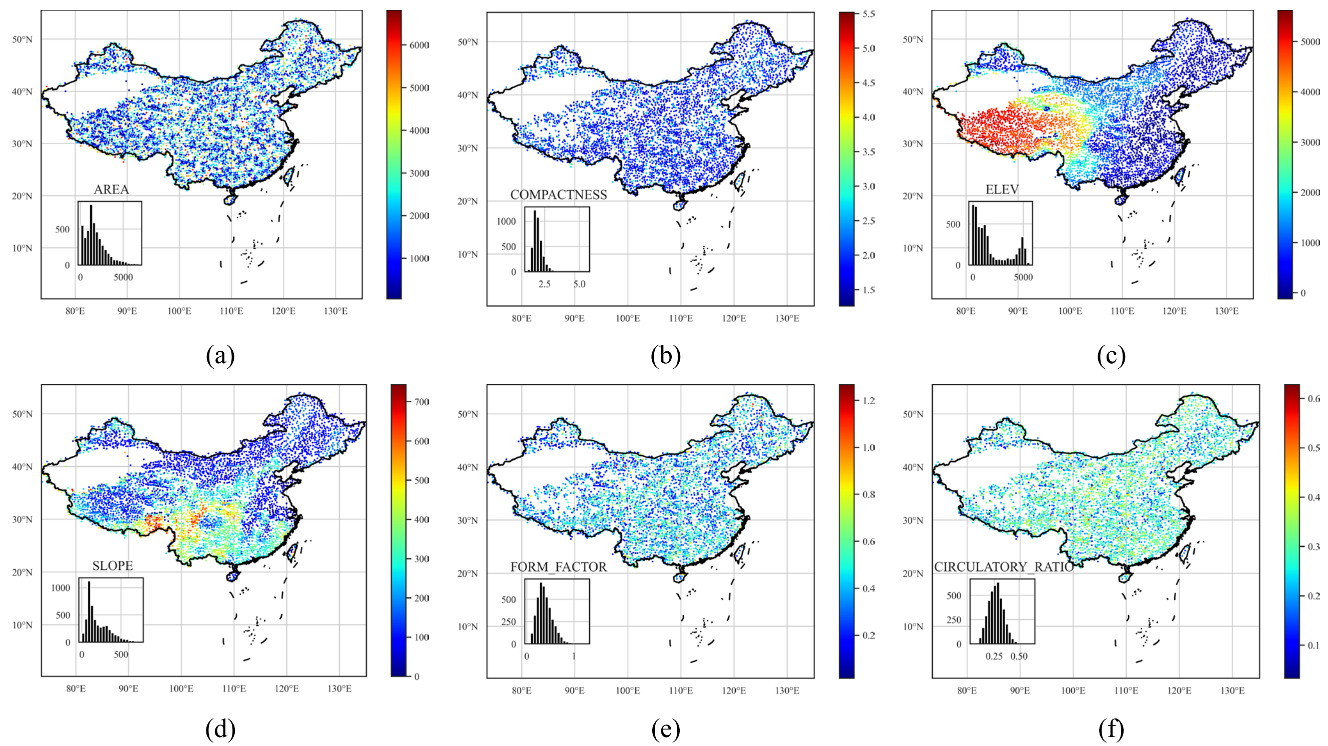
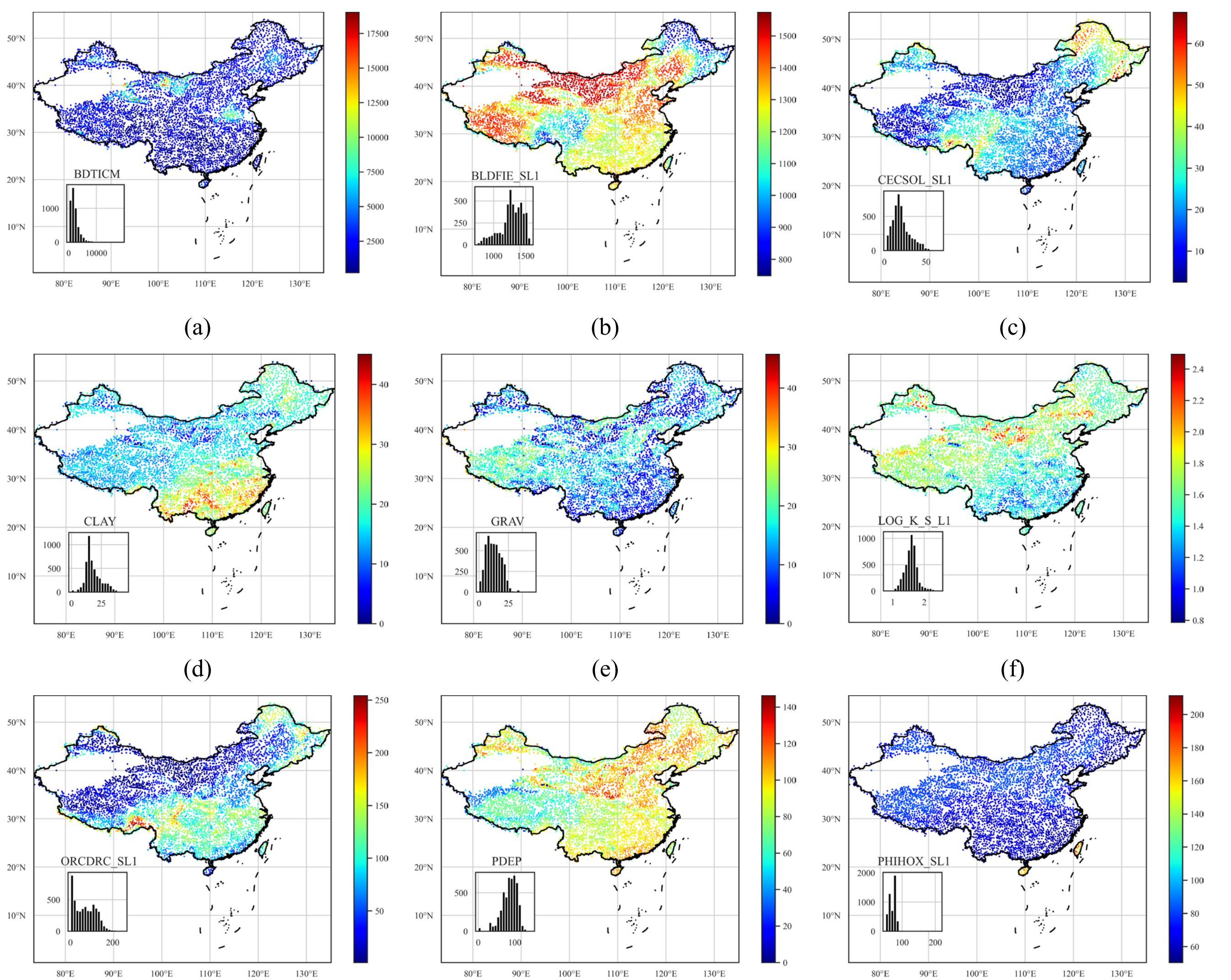


Figure 6. Distributions of topographic characteristics.

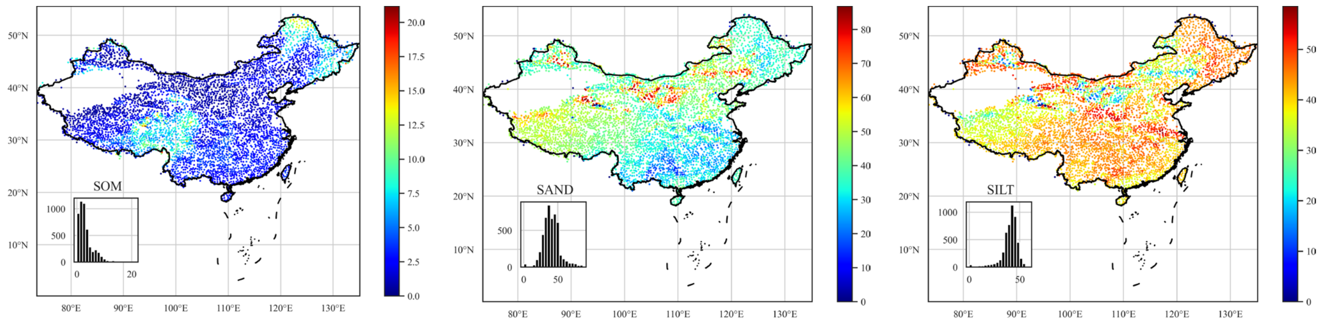
The CAMELS dataset provides two parameters (two area estimates) for describing the catchment shape. The physical characteristics of a catchment can affect the streamflow volume and the streamflow hydrograph of the catchment under a storm. To provide a complete description of the catchment shape, we computed several geometrical parameters of the catchment related to the streamflow process (Fig. 6), including catchment form factor, shape factor, compactness coefficient, circulatory ratio and the elongation ratio (Subramanya 2013). A summary of the location and topography attributes can be found in Table 3.

7 Soil

The proposed dataset has a total of 54 soil attributes (Table 3) derived from (Hengl, Mendes de Jesus et al. 2017), (Dai, Xin et al. 2019) and (Shangguan, Dai et al. 2013). Five categories of soil characteristics (pH in H2O, organic carbon content, depth to bedrock, cation-exchange capacity, and bulk density) are determined from SoilGrids. SoilGrids (Hengl, Mendes de Jesus et al. 2017) provides global predictions for soil properties including organic carbon, bulk density, cation exchange capacity (CEC), pH, soil texture fractions and coarse fragments by fusing multiple data sources including MODIS land products, SRTM DEM, climatic images and global landform and lithology maps at the 250m resolution (Fig. 7). SoilGrids made predictions based on machine learning algorithms and many covariates' layers primarily derived from remote sensing data. SoilGrids has soil characteristics for several soil depths.



(g) (h) (i)



(j) (k) (l)

Figure 7: Distributions of soil characteristics over China.

Different from CAMELS, whose reported results are obtained by a linear weighted combination of the different soil layers, and CAMELS-BR, whose products are soil characteristics at a depth of 30cm. We computed soil characteristics at all soil layers provided by SoilGrids250m.

We determined saturated water content and saturated hydraulic conductivity (Dai, Xin et al. 2019). Based on the same dataset, we also introduced the thermal conductivity of unfrozen saturated soils. Dai, Xin et al. (2019) provides a global estimation of soil hydraulic and thermal parameters using multiple Pedotransfer Functions (PTFs) based on the SoilGrids250m dataset. Based on the SoilGrids250m and GSDE (Shangguan, Dai et al. 2014) datasets, Dai, Xin et al. (2019) produced six soil layers with a spatial resolution of 30×30 arc-second. The vertical resolution of (Dai, Xin et al. 2019) is the same as the SoilGrids250m, with six intervals of 0–0.05 m, 0.05–0.15 m, 0.15–0.30 m, 0.30–0.60 m, 0.60–1.00 m, and 1.00–2.00 m. We determine and record catchment soil characteristics for all these layers. In addition, we determined seven more soil characteristics (Shangguan, Dai et al. 2013) including soil profile depth, porosity, clay/silt/sand content, rock fragment, and soil organic carbon content. Shangguan, Dai et al. (2013) provides physical and chemical attributes of soils derived from 8979 soil profiles at 30×30 arc-second resolution, the polygon linkage method was used to derive the spatial distribution of soil properties. The profile attribute database and soil map are linked under a framework avoiding uncertainty in taxon referencing.

Depth to bedrock controls many physical and chemical processes in soil. The distribution of depth to bedrock in China is characterised by (i) low in the mountainous areas, such as Yunnan province and Chongqing City; (ii) high in barren areas, e.g. North and Northwest China. The introduced soil pH value is crucial since it influences many other physical and chemical soil characteristics. The spatial variability of soil pH in China is characterised by (i) soils in southern China are acid to strongly acid; (ii) soils in northern China are natural or alkaline; (iii) soils in northeastern forested areas are also acid (pH < 7.2). Cation exchange capacity can be seen as a measure of soil fertility since it measures how much nutrient the soil can store such that it influences the growth of the vegetation. Cation exchange capacity is positively correlated with soil organic matter content and clay content, which Cation exchange capacity is generally low in sandy and silty soils. The spatial variability of Cation exchange capacity in China is characterised by (i) high in peat and forested areas in Qinghai-Tibet Plateau, central and northeast

China (ii) The Cation exchange capacity in the desert area such as the northwest is extremely low. Soil hydraulic and thermal properties are greatly affected by soil organic matter (SOM). Soil organic matter has a similar distribution to the cation exchange capacity: high in the peat and forested areas such as northeast China and low in the north and northwest.

8 Meteorological time series

Table 4: Summary table of catchment meteorological time series available in the proposed dataset

Variable	Description	Unit
prs	catchment daily averaged ground pressure	hPa
tem	catchment daily averaged temperature at 2 m above ground	°C
rhu	catchment daily averaged relative humidity	-
pre	catchment daily averaged precipitation	mm d^{-1}
evp	catchment daily averaged evaporation measured by ground instruments	mm d^{-1}
win	catchment daily averaged wind speed at 2 m above ground	m s^{-1}
ssd	catchment daily averaged sunshine duration	h d^{-1}
gst	catchment daily averaged ground surface temperature	°C
pet	catchment daily averaged potential evapotranspiration determined by Penman's equation (Appendix A)	mm d^{-1}

There have been many studies based on SURF_CLI_CHN_MUL_DAY in China (Liu, Xu et al. 2004, Xu, Gao et al. 2009, Huang, Han et al. 2016, Liu, Zheng et al. 2017), such as trend analysis of the pan evaporation (Liu, Yang et al. 2010). Still, there has not yet been a large-scale basin-oriented meteorological time series dataset in China. Researchers still need to do repeated works to extract historical meteorological data from the SURF_CLI_CHN_MUL_DAY dataset for the research. For the first time, we release a catchment scale meteorological time series dataset. The open-sourced code can generate any catchment's meteorological time series within China. The basin-oriented dataset provides meteorological time series for 4911 basins from 1990 to 2020 based on the China Meteorological Data Network. Meteorological time series includes pressure, temperature, relative humidity, precipitation, evaporation, wind speed, sunshine duration, ground surface temperature and potential evapotranspiration (Table 4).

The meteorological time series data from 1951 to 2010 is derived based on the "1951-2010 China National Ground Station Data Corrected Monthly Data File Basic Data Collection" data construction project. Other data include monthly reported data to the National Meteorological Information Centre by the provinces, and hourly and daily data uploaded by automatic ground stations in real-time. During the development of the dataset, missing data were filled by interpolating its nearest stations.

Figure 2 presents the variation of the number of sites. The start date of the recording is 1951, but because the early site distribution is sparse, we only used records from 1990 to 2020 to construct the dataset to ensure the data quality. Inverse distance weighting shows better performance than other interpolation methods. In addition, potential evapotranspiration (PET) is estimated based on Penman's Equation (Appendix A) and other meteorological variables.

9 HydroMLYR: Hydrology dataset for Machine Learning in YRB

In addition to the basin-wise static attributes provided in CCAM, we propose HydroMLYR, a hydrology dataset for machine learning research in the YRB (Fig. 1). HydroMLYR includes standardized streamflow measurements for 102 basins. The streamflow data is seven-day averaged and standardized basin-wise to have zero mean and a standard deviation of 1 (Fig. 8). The HydroMLYR dataset is proposed to support machine learning or deep learning hydrology research (e.g., neural network-based and tree-based algorithms). It can be used in two cases: (1) to develop machine learning models on the YRB or (2) when it is desirable to verify the generalization ability of a machine learning model on YRB.

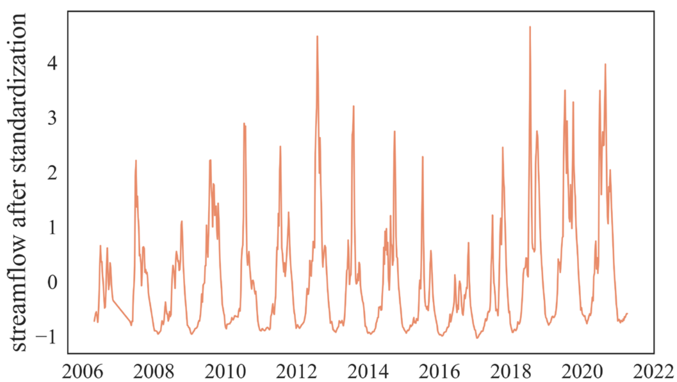


Figure 8: Examples of standardized runoff

The dataset provides 40 natural basins in the dataset which are not affected by reservoirs and dams. The selection is based on a newer version[4] of the Global Reservoirs and Dam databases (Lehner, Liermann et al. 2011) which provides the locations of reservoirs and dams globally. HydroMLYR covers 102 basins in the YRB, including basin boundary shapefiles, static attributes, and standardized streamflow measurements for each basin. The covered basins have areas ranging from 134 to 804,421 square kilometres. Therefore, modelling on a large scale of the YRB is also possible. Meteorological records in HydroMLYR introduced daily maximum and minimum for some forcing variables (Table 5).
The original streamflow observations are not continuous. The average record length is 11.3 years. Although the development of machine learning models does not necessarily require the data to be continuous, we separately provide continuous streamflow observations with an average record length of 8.3 years.

[4] http://globaldamwatch.org/data/#core_global

Table 5: Meteorological variables provided in HydroMLYR

Attribute name	Description	Unit
evp	catchment daily averaged evaporation (observations)	mm d^{-1}
gst_mean	catchment daily averaged ground surface temperature	°C
gst_min	catchment daily minimum ground surface temperature	°C
gst_max	catchment daily maximum ground surface temperature	°C
pre	catchment daily averaged precipitation	mm d^{-1}
prs_mean	catchment daily averaged ground surface pressure	hPa
prs_max	catchment daily maximum ground surface pressure	hPa
prs_min	catchment daily minimum ground surface pressure	hPa
rhu	catchment daily averaged relative humidity	-
ssd	catchment daily averaged sunshine duration	h
tem_mean	catchment daily averaged temperature	°C
tem_min	catchment daily minimum temperature	°C
tem_max	catchment daily maximum temperature	°C
win_max	catchment daily maximum wind speed	m s^{-1}
win_mean	catchment daily averaged wind speed	m s^{-1}

10 Data and code availability

The proposed dataset is freely available at http://doi.org/10.5281/zenodo.5137288. The files provided are (i) several separate files containing 120+ catchments attributes, (ii) the daily meteorological time series in a zip file, (iii) the catchment boundaries used to compute the attributes and extract the time series, (iv) the HydroMLYR dataset, (v) an attribute description file and (v) a readme file.

11 Conclusion

The CCAM dataset proposed in this paper provides a novel dataset for hydrological research in China. All basins delaminated from the DEM are studied, covering entire China. The dataset includes daily meteorological forcing time-series data including precipitation, temperature, potential evapotranspiration, wind, ground surface temperature, pressure, humidity, sunshine duration and derived potential evapotranspiration of 4911 catchments. The proposed time series dataset is derived based on the quality-controlled SURF_CLI_CHN_MUL_DAY dataset. CCAM includes 120+ catchment attributes, including soil, land cover, geology, climate indices and topography for each catchment. We produced a series of maps depicting the catchment attributes distributions in China. These maps present regional changes of various features; we also estimate the relationships between them based on Kendall's correlation. Integrating multiple data sources into one dataset at a catchment scale simplifies the data compilation process in research. CCAM can help test hypotheses and formulate valid conclusions under various conditions, not just limited to a few specific locations and help explore how different basin characteristics influence hydrological behaviours, learn the migration of hydrological behaviours between different basins, and develop general frameworks for large-scale model evaluation and benchmarking in China. A limitation of the study is the lack of estimation of the uncertainty of the meteorological time series. An alternative is to evaluate the uncertainty of the basin-wise meteorological data based on multiple independent data sources, but there are few data that provide as many data types as SURF_CLI_CHN_MUL_DAY. Hence, it poses a challenge for evaluating the uncertainty of these eight meteorological variables, which is left for future studies.

Appendix A: Modified Penman's equation

Penman's equation (Subramanya 2013), incorporating some modifications to the original formula, is:

$$PET = \frac{AH_n + E_a\gamma}{A + \gamma}$$

where PET is the daily potential evapotranspiration in mm per day; A is the slope of the saturation vapour pressure (ew) vs temperature (t) curve at the mean air temperature, in mm of mercury per Celsius; Hn is the net radiation in mm of evaporable water per day; Ea is a parameter including wind speed and saturation deficit; γ is the psychrometric constant = 0.49 mm of mercury per Celsius.

The relationship between ew and t is defined as:

$$e_w = 4.584 \exp\left(\frac{17.27t}{237.3 + t}\right)$$

The following equation estimates the net radiation:

$$H_n = H_a(1 - r)\left(a + b\frac{n}{N}\right) - \sigma T_a^4\left(0.56 - 0.092\sqrt{e_a}\right)\left(0.10 + 0.90\frac{n}{N}\right)$$

where Ha is the incident solar radiation outside the atmosphere on a horizontal surface, expressed in mm of evaporable water per day (a function of the latitude and period of the year as indicated in Table A1); a is a constant depending upon the latitude ϕ and is given by $a = 0.29\ cos\ \phi$; b is a constant = 0.52; n is the sunshine duration in hours; N is the maximum possible hours of bright sunshine (a function of latitude, see Table A2); r is the reflection coefficient; σ is the Stefan-Boltzman constant = 2.01×10^{-9} mm/day; Ta is the mean air temperature in degrees kelvin; ea is the actual mean vapour pressure in the air in mm of mercury.

Table A1: Mean Monthly Solar Radiation, Ha in mm of Evaporable Water/Day

North latitude	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
0°	14.5	15.0	15.2	14.7	13.9	13.4	13.5	14.2	14.9	15.0	14.6	14.3
10°	12.8	13.9	14.8	15.2	15.0	14.8	14.8	15.0	14.9	14.1	13.1	12.4
20°	10.8	12.3	13.9	15.2	15.7	15.8	15.7	15.3	14.4	12.9	11.2	10.3
30°	8.5	10.5	12.7	14.8	16.0	16.5	16.2	15.3	13.5	11.3	9.1	7.9
40°	6.0	8.3	11.0	13.9	15.9	16.7	16.3	14.8	12.2	9.3	6.7	5.4
50°	3.6	5.9	9.1	12.7	15.4	16.7	16.1	13.9	10.5	7.1	4.3	3.0

The parameter Ea is estimated as:

$$E_a = 0.35\left(1 + \frac{u_2}{160}\right)(e_w - e_a)$$

where $u2$ is the wind speed at $2m$ above ground in km/day; ew is the saturation vapour pressure at mean air temperature in mm of mercury; ea is the actual vapour pressure.

Table A2: Mean Monthly Values of Possible Sunshine Hours, N

North latitude	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
0°	12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1
10°	11.6	11.8	12.1	12.4	12.6	12.7	12.6	12.4	12.9	11.9	11.7	11.5
20°	11.1	11.5	12.0	12.6	13.1	13.3	13.2	12.8	12.3	11.7	11.2	10.9
30°	10.4	11.1	12.0	12.9	13.7	14.1	13.9	13.2	12.4	11.5	10.6	10.2
40°	9.6	10.7	11.9	13.2	14.4	15.0	14.7	13.8	12.5	11.2	10.0	9.4
50°	8.6	10.1	11.8	13.8	15.4	16.4	16.0	14.5	12.7	10.8	9.1	8.1

Appendix B: Correlation analysis of catchment attributes

To explore the potential connections between various types of watershed attributes, we did correlation analysis using the Kendall rank correlation coefficient (Kendall 1938). Kendall rank correlation coefficient is a measure of rank correlation: the similarity of the sort order of the two sets of data. Kendall correlation will be high if the orderings of the observations of two variables are similar. Kendall correlation avoids the assumption of linear relationship and that the distribution should be normal and continuous (e.g., Pearson correlation). When the relationship is not exactly linear, using Pearson correlation will miss out on information that Kendall could capture. Table B1 shows the top five most relevant attributes for each attribute. The analysis result shows that the correlations between variables are in line with general understanding, justifying the rationality of the dataset, to name a few:

(1) Subsurface permeability and porosity are most correlated with geological attributes.

(2) LAI and NDVI are most positively correlated with each other but most negatively correlated with the fraction of barren land cover.

(3) Urban and built ups are most positively correlated with population density.

(4) In China, the savanna is mainly distributed in the southern coastal areas, resulting in that it is most positively correlated with mean precipitation.

(5) Sand is most positively correlated with the saturated hydraulic conductivity while the clay is strongly negatively correlated with.

Table B1: The top five most relevant characteristics for each attribute (different soil layers for the same attribute are excluded, e.g., phihox_sl2 is not included in the top five most relevant attributes of phihox_sl1 though they are highly correlated)

Attribute	1st	2nd	3rd	4th	5th
high_prec_freq	root_depth_50(-0.196)	grassland(0.175)	root_depth_99(-0.171)	som(0.136)	tksatu_l1(-0.133)
high_prec_dur	theta_s_l6(-0.277)	theta_s_l5(-0.234)	p_seasonality(0.233)	elev(0.211)	theta_s_l4(-0.201)
low_prec_freq	pre_mean(-0.766)	aridity(0.745)	ssd_mean(0.652)	rhu_mean(-0.627)	phihox_sl7(0.588)
low_prec_dur	aridity(0.78)	pre_mean(-0.768)	ssd_mean(0.731)	rhu_mean(-0.709)	phihox_sl7(0.579)
frac_snow_daily	gst_mean(-0.802)	tem_mean(-0.792)	lat(0.575)	evergreen_broadleaf_tree(-0.512)	pre_mean(-0.436)
prs_mean	elev(-0.678)	lon(0.552)	rhu_mean(0.432)	urban_and_built-up_land(0.427)	barren(-0.41)
pre_mean	aridity(-0.913)	low_prec_dur(-0.768)	low_prec_freq(-0.766)	ssd_mean(-0.723)	rhu_mean(0.712)
evp_mean	aridity(0.643)	ndvi_mean(-0.632)	rhu_mean(-0.617)	ssd_mean(0.598)	lai_dif(-0.593)
gst_mean	tem_mean(0.924)	frac_snow_daily(-0.802)	lat(-0.512)	evergreen_broadleaf_tree(0.507)	pet_mean(0.442)
rhu_mean	aridity(-0.751)	ssd_mean(-0.746)	pre_mean(0.712)	low_prec_dur(-0.709)	low_prec_freq(-0.627)
pet_mean	cecsol_sl2(-0.451)	gst_mean(0.442)	cecsol_sl3(-0.441)	cecsol_sl1(-0.422)	cecsol_sl4(-0.42)
ssd_mean	aridity(0.753)	rhu_mean(-0.746)	low_prec_dur(0.731)	pre_mean(-0.723)	low_prec_freq(0.652)

win_mean	ssd_mean(0.426)	woody_savanna(-0.393)	tem_mean(-0.379)	gst_mean(-0.377)	mixed_forest(-0.363)
tem_mean	gst_mean(0.924)	frac_snow_daily(-0.792)	evergreen_broadleaf_tree(0.493)	pop_dnsty(0.475)	lat(-0.474)
p_seasonality	rhu_mean(-0.421)	tem_mean(-0.397)	gst_mean(-0.393)	ssd_mean(0.393)	low_prec_dur(0.375)
aridity	pre_mean(-0.913)	low_prec_dur(0.78)	ssd_mean(0.753)	rhu_mean(-0.751)	low_prec_freq(0.745)
slope	lat(-0.374)	bdticm(-0.348)	win_mean(-0.341)	mixed_forest(0.341)	evergreen_needleleaf_tree(0.327)
lon	elev(-0.585)	prs_mean(0.552)	evp_mean(-0.5)	barren(-0.482)	ndvi_mean(0.47)
elev	prs_mean(-0.678)	lon(-0.585)	urban_and_built-up_land(-0.485)	pop_dnsty(-0.481)	cropland(-0.456)
lat	frac_snow_daily(0.575)	evergreen_broadleaf_tree(-0.548)	gst_mean(-0.512)	tem_mean(-0.474)	low_prec_freq(0.437)
pop	urban_and_built-up_land(0.618)	cropland(0.519)	aridity(-0.511)	pre_mean(0.505)	rhu_mean(0.492)
pop_dnsty	urban_and_built-up_land(0.639)	aridity(-0.538)	cropland(0.533)	pre_mean(0.533)	ssd_mean(-0.521)
length	area(0.684)	form_factor(-0.398)	shape_factor(0.398)	elongation_ratio(-0.398)	compactness_coefficient(0.363)
area	length(0.684)	pop(0.23)	pa(0.194)	circulatory_ratio(-0.187)	compactness_coefficient(0.187)
form_factor	elongation_ratio(1.0)	shape_factor(-1.0)	circulatory_ratio(0.435)	compactness_coefficient(-0.435)	length(-0.398)
shape_factor	elongation_ratio(-1.0)	form_factor(-1.0)	circulatory_ratio(-0.435)	compactness_coefficient(0.435)	length(0.398)
compactness_coefficient	circulatory_ratio(-1.0)	elongation_ratio(-0.435)	shape_factor(0.435)	form_factor(-0.435)	length(0.363)
circulatory_ratio	compactness_coefficient(-1.0)	elongation_ratio(0.435)	shape_factor(-0.435)	form_factor(0.435)	length(-0.363)
elongation_ratio	shape_factor(-1.0)	form_factor(1.0)	circulatory_ratio(0.435)	compactness_coefficient(-0.435)	length(-0.398)

lai_dif	ndvi_mean(0.808)	barren(-0.642)	aridity(-0.638)	pre_mean(0.609)	woody_savanna(0.607)
lai_max	ndvi_mean(0.779)	barren(-0.614)	aridity(-0.613)	woody_savanna(0.612)	phihox_sl2(-0.602)
ndvi_mean	lai_dif(0.808)	lai_max(0.779)	barren(-0.677)	evp_mean(-0.632)	aridity(-0.607)
root_depth_50	grassland(-0.485)	pet_mean(0.232)	barren(0.212)	high_prec_freq(-0.196)	pdep(-0.176)
root_depth_99	grassland(-0.339)	barren(0.337)	cropland(-0.336)	pdep(-0.284)	lon(-0.283)
evergreen_needleleaf_tree	mixed_forest(0.572)	woody_savanna(0.481)	phihox_sl7(-0.416)	phihox_sl6(-0.411)	phihox_sl5(-0.409)
evergreen_broadleaf_tree	lat(-0.548)	phihox_sl7(-0.538)	phihox_sl6(-0.529)	phihox_sl5(-0.522)	pre_mean(0.512)
deciduous_needleleaf_tree	cecsol_sl1(0.274)	bldfie_sl1(-0.274)	cecsol_sl2(0.272)	orcdrc_sl2(0.27)	cecsol_sl3(0.262)
deciduous_broadleaf_tree	mixed_forest(0.604)	woody_savanna(0.568)	ndvi_mean(0.524)	lai_max(0.5)	lai_dif(0.497)
mixed_forest	woody_savanna(0.713)	deciduous_broadleaf_tree(0.604)	evergreen_needleleaf_tree(0.572)	phihox_sl7(-0.565)	phihox_sl6(-0.563)
closed_shrubland	deciduous_broadleaf_tree(0.217)	savanna(0.16)	mixed_forest(0.158)	tksatu_l4(-0.153)	theta_s_l2(-0.142)
open_shrubland	high_prec_dur(0.179)	rhu_mean(-0.174)	elev(0.17)	ssd_mean(0.17)	prs_mean(-0.165)
woody_savanna	mixed_forest(0.713)	phihox_sl7(-0.628)	phihox_sl4(-0.628)	phihox_sl3(-0.627)	phihox_sl6(-0.627)
savanna	pre_mean(0.606)	cropland_natural_vegetaion(0.605)	woody_savanna(0.604)	aridity(-0.602)	ssd_mean(-0.591)
grassland	root_depth_50(-0.485)	cropland_natural_vegetaion(-0.363)	tem_mean(-0.344)	gst_mean(-0.344)	root_depth_99(-0.339)
permanent_wetland	water_bodies(0.469)	savanna(0.363)	urban_and_built-up_land(0.347)	pre_mean(0.343)	pop(0.343)
cropland	urban_and_built-up_land(0.546)	pop_dnsty(0.533)	pop(0.519)	elev(-0.456)	lon(0.417)

urban_and_built-up_land	pop_dnsty(0.639)	pop(0.618)	cropland(0.546)	elev(-0.485)	cropland_natural_vegetaion(0.428)
cropland_natural_vegetaion	savanna(0.605)	rhu_mean(0.546)	aridity(-0.523)	ssd_mean(-0.52)	pre_mean(0.51)
snow_and_ice	ig(0.431)	barren(0.379)	lon(-0.373)	elev(0.369)	pdep(-0.354)
barren	ndvi_mean(-0.677)	lai_dif(-0.642)	lai_max(-0.614)	aridity(0.581)	evp_mean(0.574)
water_bodies	permanent_wetland(0.469)	wb(0.39)	cropland_natural_vegetaion(0.17)	urban_and_built-up_land(0.158)	elev(-0.154)
geol_permeability	sm(-0.345)	su(0.326)	ss(-0.316)	bdticm(0.228)	pdep(0.161)
geol_porosity	su(0.455)	pa(-0.417)	woody_savanna(-0.323)	phihox_sl3(0.315)	phihox_sl4(0.314)
ig	snow_and_ice(0.431)	elev(0.194)	theta_s_l2(-0.185)	pdep(-0.184)	theta_s_l3(-0.182)
pa	geol_porosity(-0.417)	mt(0.3)	pi(0.295)	va(0.271)	vi(0.246)
sc	geol_porosity(-0.285)	lat(-0.264)	bdticm(-0.26)	slope(0.246)	mixed_forest(0.231)
su	bdticm(0.52)	geol_porosity(0.455)	woody_savanna(-0.349)	geol_permeability (0.326)	phihox_sl7(0.326)
sm	geol_permeability(-0.345)	su(-0.283)	bdticm(-0.228)	cropland(-0.199)	elev(0.194)
vi	pa(0.246)	pi(0.203)	va(0.171)	geol_porosity(-0.169)	deciduous_broadleaf_tree(0.166)
mt	pa(0.3)	geol_porosity(-0.286)	pi(0.199)	deciduous_broadleaf_tree(0.187)	area(0.18)
ss	geol_permeability(-0.316)	su(-0.17)	bdticm(-0.136)	evergreen_needleleaf_tree(0.106)	tksatu_l6(-0.096)
pi	pa(0.295)	vi(0.203)	mt(0.199)	geol_porosity(-0.183)	va(0.172)
va	pa(0.271)	geol_porosity(-0.219)	vb(0.21)	deciduous_needleleaf_tree(0.186)	pi(0.172)

wb	water_bodies(0.39)	permanent_wetland(0.264)	bldfie_sl4(0.148)	bldfie_sl5(0.147)	urban_and_built-up_land(0.138)
pb	mt(0.176)	pa(0.132)	theta_s_l5(-0.128)	area(0.127)	length(0.123)
vb	va(0.21)	geol_porosity(-0.171)	vi(0.165)	cecsol_sl7(0.161)	cecsol_sl6(0.157)
nd	barren(0.154)	aridity(0.146)	pre_mean(-0.144)	lai_dif(-0.141)	snow_and_ice(0.141)
py	phihox_sl1(-0.237)	phihox_sl2(-0.233)	phihox_sl3(-0.233)	phihox_sl4(-0.23)	woody_savanna(0.227)
ev	barren(0.036)	orcdrc_sl5(-0.035)	orcdrc_sl4(-0.035)	cecsol_sl3(-0.034)	orcdrc_sl7(-0.034)
tksatu_l1	grav(-0.346)	som(-0.344)	bldfie_sl3(0.298)	bldfie_sl1(0.295)	bldfie_sl2(0.291)
tksatu_l2	som(-0.365)	bldfie_sl3(0.326)	bldfie_sl1(0.326)	bldfie_sl2(0.323)	grav(-0.308)
tksatu_l3	som(-0.344)	bldfie_sl2(0.328)	bldfie_sl1(0.325)	bldfie_sl3(0.324)	bldfie_sl4(0.308)
tksatu_l4	bldfie_sl2(0.398)	som(-0.397)	bldfie_sl1(0.388)	bldfie_sl3(0.384)	bldfie_sl4(0.358)
tksatu_l5	bldfie_sl3(0.386)	bldfie_sl2(0.376)	som(-0.369)	bldfie_sl4(0.364)	bldfie_sl1(0.358)
tksatu_l6	bldfie_sl3(0.366)	som(-0.362)	bdticm(0.36)	bldfie_sl2(0.343)	bldfie_sl7(0.338)
log_k_s_l1	sand(0.71)	clay(-0.59)	savanna(-0.441)	silt(-0.436)	rhu_mean(-0.423)
log_k_s_l2	sand(0.709)	clay(-0.578)	savanna(-0.452)	phihox_sl7(0.438)	silt(-0.433)
log_k_s_l3	sand(0.682)	clay(-0.592)	savanna(-0.448)	phihox_sl7(0.442)	phihox_sl6(0.435)
log_k_s_l4	sand(0.612)	clay(-0.603)	savanna(-0.49)	pre_mean(-0.489)	phihox_sl7(0.485)
log_k_s_l5	clay(-0.561)	sand(0.555)	phihox_sl7(0.506)	savanna(-0.501)	phihox_sl6(0.501)
log_k_s_l6	clay(-0.563)	pre_mean(-0.555)	aridity(0.548)	phihox_sl7(0.534)	phihox_sl6(0.532)
theta_s_l1	grav(-0.582)	clay(0.325)	sand(-0.315)	elev(-0.314)	pdep(0.311)
theta_s_l2	grav(-0.585)	pdep(0.377)	elev(-0.366)	clay(0.35)	sand(-0.326)
theta_s_l3	grav(-0.522)	pdep(0.42)	elev(-0.414)	prs_mean(0.365)	clay(0.359)
theta_s_l4	grav(-0.515)	pdep(0.463)	elev(-0.412)	prs_mean(0.349)	lon(0.328)
theta_s_l5	grav(-0.433)	elev(-0.401)	pdep(0.376)	sand(-0.349)	rhu_mean(0.331)
theta_s_l6	evergreen_broadleaf_tree(0.372)	grav(-0.357)	elev(-0.344)	sand(-0.343)	tem_mean(0.337)
orcdrc_sl7	bldfie_sl4(-0.581)	bldfie_sl5(-0.572)	bldfie_sl6(-0.548)	bldfie_sl3(-0.535)	bldfie_sl7(-0.523)

orcdrc_sl3	bldfie_sl3(-0.738)	bldfie_sl2(-0.728)	bldfie_sl1(-0.701)	bldfie_sl4(-0.691)	bldfie_sl5(-0.621)
orcdrc_sl4	bldfie_sl3(-0.702)	bldfie_sl2(-0.682)	bldfie_sl4(-0.676)	bldfie_sl1(-0.657)	bldfie_sl5(-0.614)
orcdrc_sl5	bldfie_sl4(-0.641)	bldfie_sl3(-0.636)	bldfie_sl2(-0.611)	bldfie_sl5(-0.6)	bldfie_sl1(-0.592)
orcdrc_sl6	bldfie_sl4(-0.584)	bldfie_sl5(-0.567)	bldfie_sl6(-0.556)	bldfie_sl3(-0.552)	bldfie_sl7(-0.534)
orcdrc_sl2	bldfie_sl2(-0.787)	bldfie_sl1(-0.769)	bldfie_sl3(-0.749)	bldfie_sl4(-0.68)	cecsol_sl1(0.629)
orcdrc_sl1	phihox_sl2(-0.599)	phihox_sl3(-0.594)	phihox_sl4(-0.591)	phihox_sl5(-0.586)	phihox_sl6(-0.585)
phihox_sl7	woody_savanna(-0.628)	pre_mean(-0.598)	aridity(0.592)	low_prec_freq(0.588)	orcdrc_sl1(-0.583)
phihox_sl6	woody_savanna(-0.627)	pre_mean(-0.594)	aridity(0.59)	lai_max(-0.587)	orcdrc_sl1(-0.585)
phihox_sl5	woody_savanna(-0.626)	lai_max(-0.593)	pre_mean(-0.592)	aridity(0.589)	orcdrc_sl1(-0.586)
phihox_sl4	woody_savanna(-0.628)	lai_max(-0.599)	orcdrc_sl1(-0.591)	lai_dif(-0.578)	pre_mean(-0.576)
phihox_sl3	woody_savanna(-0.627)	lai_max(-0.595)	orcdrc_sl1(-0.594)	lai_dif(-0.576)	pre_mean(-0.568)
phihox_sl2	woody_savanna(-0.627)	lai_max(-0.602)	orcdrc_sl1(-0.599)	lai_dif(-0.583)	low_prec_freq(0.569)
phihox_sl1	woody_savanna(-0.601)	lai_max(-0.586)	orcdrc_sl1(-0.584)	lai_dif(-0.565)	bldfie_sl2(0.55)
bldfie_sl7	orcdrc_sl5(-0.547)	orcdrc_sl4(-0.546)	orcdrc_sl3(-0.543)	orcdrc_sl6(-0.534)	orcdrc_sl7(-0.523)
bldfie_sl6	orcdrc_sl5(-0.559)	orcdrc_sl6(-0.556)	orcdrc_sl4(-0.553)	orcdrc_sl7(-0.548)	orcdrc_sl3(-0.547)
bldfie_sl5	orcdrc_sl3(-0.621)	orcdrc_sl4(-0.614)	orcdrc_sl5(-0.6)	orcdrc_sl2(-0.597)	orcdrc_sl7(-0.572)

bldfie_sl4	orcdrc_sl3(-0.691)	orcdrc_sl2(-0.68)	orcdrc_sl4(-0.676)	orcdrc_sl5(-0.641)	orcdrc_sl6(-0.584)
bldfie_sl1	orcdrc_sl2(-0.769)	orcdrc_sl3(-0.701)	cecsol_sl1(-0.686)	orcdrc_sl4(-0.657)	som(-0.606)
bldfie_sl3	orcdrc_sl2(-0.749)	orcdrc_sl3(-0.738)	orcdrc_sl4(-0.702)	orcdrc_sl5(-0.636)	som(-0.633)
bldfie_sl2	orcdrc_sl2(-0.787)	orcdrc_sl3(-0.728)	orcdrc_sl4(-0.682)	cecsol_sl1(-0.671)	som(-0.651)
cecsol_sl1	bldfie_sl1(-0.686)	bldfie_sl2(-0.671)	orcdrc_sl2(0.629)	bldfie_sl3(-0.598)	orcdrc_sl3(0.579)
cecsol_sl2	bldfie_sl1(-0.579)	bldfie_sl2(-0.566)	orcdrc_sl2(0.553)	orcdrc_sl3(0.523)	bldfie_sl3(-0.515)
cecsol_sl5	bldfie_sl1(-0.445)	bldfie_sl2(-0.429)	orcdrc_sl2(0.412)	orcdrc_sl3(0.393)	pet_mean(-0.392)
cecsol_sl4	bldfie_sl1(-0.472)	bldfie_sl2(-0.459)	orcdrc_sl2(0.447)	orcdrc_sl3(0.43)	orcdrc_sl5(0.424)
cecsol_sl3	bldfie_sl1(-0.532)	bldfie_sl2(-0.52)	orcdrc_sl2(0.508)	orcdrc_sl3(0.49)	orcdrc_sl4(0.478)
cecsol_sl7	bldfie_sl1(-0.413)	bldfie_sl2(-0.396)	orcdrc_sl2(0.38)	pet_mean(-0.374)	orcdrc_sl3(0.362)
cecsol_sl6	bldfie_sl1(-0.409)	bldfie_sl2(-0.393)	orcdrc_sl2(0.378)	pet_mean(-0.373)	orcdrc_sl3(0.36)
bdticm	su(0.52)	woody_savanna(-0.412)	low_prec_freq(0.382)	phihox_sl7(0.378)	mixed_forest(-0.374)
pdep	theta_s_l4(0.463)	elev(-0.436)	grav(-0.424)	theta_s_l3(0.42)	lon(0.4)
por	som(0.363)	bldfie_sl1(-0.335)	phihox_sl1(-0.329)	phihox_sl3(-0.328)	phihox_sl2(-0.328)
clay	sand(-0.67)	log_k_s_l4(-0.603)	log_k_s_l3(-0.592)	log_k_s_l1(-0.59)	log_k_s_l2(-0.578)
sand	log_k_s_l1(0.71)	log_k_s_l2(0.709)	log_k_s_l3(0.682)	clay(-0.67)	log_k_s_l4(0.612)
silt	sand(-0.573)	log_k_s_l1(-0.436)	log_k_s_l2(-0.433)	log_k_s_l3(-0.4)	log_k_s_l4(-0.316)

grav	theta_s_l2(-0.585)	theta_s_l1(-0.582)	theta_s_l3(-0.522)	theta_s_l4(-0.515)	theta_s_l5(-0.433)
som	bldfie_sl2(-0.651)	bldfie_sl3(-0.633)	bldfie_sl1(-0.606)	orcdrc_sl2(0.599)	orcdrc_sl3(0.576)
high_prec_freq	root_depth_50(-0.196)	grassland(0.175)	root_depth_99(-0.171)	som(0.136)	tksatu_l1(-0.133)
high_prec_dur	theta_s_l6(-0.277)	theta_s_l5(-0.234)	p_seasonality(0.233)	elev(0.211)	theta_s_l4(-0.201)
low_prec_freq	pre_mean(-0.766)	aridity(0.745)	ssd_mean(0.652)	rhu_mean(-0.627)	phihox_sl7(0.588)

Appendix C: Data sources and data processing

The program to generate the data set is mainly written in Python. The rasterio[5] library is used to extract from the raster for the given basin boundary, reproject and merge rasters; The shapely[6] library is used to calculate the geometry; The pyproj[7] library is used for coordinate system conversions; The richdem[8] library is used to calculate slope; The netCDF4[9] and xarray[10] library is used to read the netCDF files; The pyshp[11] library is used to handle shapefiles; The gdal[12] command-line programs are used for data format conversions; The Python multiprocessing[13] library is used for multi-threaded data processing such as the calculation of meteorological time series; The interpolation program is written based on SciPy and NumPy. In addition, the calculation of the catchment boundary uses ArcPy[14]. However, ArcPy is not open sourced. The SURF_CLI_CHN_MUL_DAY dataset can be downloaded from https://data.cma.cn/data/cdcdetail/dataCode/SURF_CLI_CHN_MUL_DAY.html. It is freely available to global researchers but registration is required. The GDBD dataset can be downloaded at https://www.cger.nies.go.jp/db/gdbd/gdbd_index_e.html. ASTER GDEM dataset can be downloaded at: https://asterweb.jpl.nasa.gov/gdem.asp. GLHYMPS dataset can be downloaded at: https://dataverse.scholarsportal.info/dataset.xhtml?persistentId=doi:10.5683/SP2/DLGXYO; MODIS MCD12Q1 can be

[5] https://rasterio.readthedocs.io/en/latest/

[6] https://shapely.readthedocs.io/en/stable/manual.html

[7] https://pyproj4.github.io/pyproj/stable/

[8] https://richdem.readthedocs.io/en/latest/

[9] https://unidata.github.io/netcdf4-python/

[10] http://xarray.pydata.org/en/stable/

[11] https://pypi.org/project/pyshp/

[12] https://gdal.org/api/python.html

[13] https://docs.python.org/3/library/multiprocessing.html

[14] https://pro.arcgis.com/zh-cn/pro-app/latest/arcpy/get-started/what-is-arcpy-.htm

obtained from: https://lpdaac.usgs.gov/products/mcd12q1v006/; MODIS MCD15A3 can be obtained from: https://lpdaac.usgs.gov/products/mcd15a3hv006/; Soil hydraulic and thermal properties can be downloaded after registration: http://globalchange.bnu.edu.cn/research/soil5.jsp; Soil properties data can be downloaded after registration: http://globalchange.bnu.edu.cn/research/soil2; SoilGrids250m data download links: https://files.isric.org/soilgrids/former/2017-03-10/data/ with a list of descriptions: https://github.com/ISRICWorldSoil/SoilGrids250m/blob/master/grids/models/META_GEOTIFF_1B.csv.

Appendix D: Basin boundaries

This section briefly introduces how the basin boundaries are derived. The basin boundaries data used in this research are obtained from the GBDB (Masutomi, Inui et al. 2009) dataset. The GDBD dataset first distinguishing sinks caused by DEM errors, then the stream burning (Maidment 1996), and ridge fencing methods are used to modify the seeded DEM, then basin boundaries are produced with standardized procedures (Jenson, Domingue et al. 1988, Maidment and Morehouse 2002). Then the gauging station data from the GRDC (Center 2005) dataset is used to calibrate the derived basin boundaries. The derived basin areas were compared with the observed basin areas, and they showed a high degree of consistency with the observed basin data.

Appendix E: Guidelines for generating basin attributes for any basin

The published code[15] supports the automation of the calculation of the attributes for any given river basin and the generation of statistics files. In general, the user only needs to prepare the source data and ensure that the code environment is installed correctly, and then the user can run the code to calculate all attributes for the given river basin. The following describes the steps to generate data for any given watershed.

Prepare source data

In this step, the user needs to download the source data and place it in the corresponding location (Table D1). The code supports the calculation of meteorological time series based on the SURF_CLI_CHN_MUL_DAY data set. If the basin the user need to calculate is not in China, then the user needs to format the collected meteorological time series into the same format as the time series generated by the code. A sample file is available in the GitHub library.

[15] https://github.com/haozhen315/CCAM-China-Catchment-Attributes-and-Meteorology-dataset

Table D1: Instructions for preparing data sources

Data source	Download link	Example	Note
ASTER GDEM	https://search.earthdata.nasa.gov/search/ https://www.jspacesystems.or.jp/ersdac/GDEM/E/	./data/dems/ *.tif	
GLHYMPS	https://dataverse.scholarsportal.info/dataset.xhtml?persistentId=doi:10.5683/SP2/DLGXYO (using source data requires merging multiple small pieces to a single TIFF) https://1drv.ms/u/s!AqzR0fLyn9KKspF6HAAuXU9Twkkz1Q?e=QCPFAm (our processed file) https://1drv.ms/u/s!AqzR0fLyn9KKspF70EPmDubS5V2qTQ?e=Rbybwa (our processed file)	./data/processed_permeability.tif ./data/processed_porosity.tif	
GLiM	https://csdms.colorado.edu/wiki/Data:GLiM https://1drv.ms/u/s!AqzR0fLyn9KKspF5Vktb-zlmd_Ctxg?e=G6fOuh (our processed file)	./data/processed_glim.py	
MCD12Q1	https://lpdaac.usgs.gov/products/mcd12q1v006/ https://1drv.ms/u/s!AqzR0fLyn9KKspF4xxbe0xM7qJNzkA?e=vyFcFj (our processed file)	./data/processed_igbp.tif	
MCD15A3	https://lpdaac.usgs.gov/products/mcd15a3hv006/	./data/MCD15A3/MCD15A3H.A2002185.h22v04.006.2015149102803.hdf	

MOD13Q1	https://lpdaac.usgs.gov/products/mod13q1v006/	./data/MOD13Q1/MOD13Q1.A2002186.h22v04.006.2015149102803.hdf	
Soil	http://globalchange.bnu.edu.cn/research/soil5.jsp	./data/soil_souce_data/binary/log_k_s_l1	
Soil	https://files.isric.org/soilgrids/former/2017-03-10/data/	./data/soil_souce_data/tif/BDTICM_M_250m_ll.tif	Description: https://github.com/ISRICWorldSoil/SoilGrids250m/blob/master/grids/models/META_GEOTIFF_1B.csv
Soil	http://globalchange.bnu.edu.cn/research/soil2	./data/soil_souce_data/tif/SA.nc	
SURF_CLI_CHN_MUL_DAY	https://data.cma.cn/data/cdcdetail/dataCode/SURF_CLI_CHN_MUL_DAY.html	./data/SURF_CLI_CHN_MUL_DAY/Data/EVP/SURF_CLI_CHN_MUL_DAY-EVP-13240-195101.TXT	If basin boundary is outside China, format and place the collected time series data in ./output/catchment_meteorological
Root depth	https://github.com/haozhen315/CCAM-China-Catchment-Attributes-and-Meteorology-dataset/blob/main/data/root_depth_calculated.txt	./data/root_depth_calculated.txt	Calculated root depth of each land type according to (Zeng 2001).
GLiM name mapping	https://github.com/haozhen315/CCAM-China-Catchment-Attributes-and-Meteorology-dataset/blob/main/data/glim_name_short_long.txt https://github.com/haozhen315/CCAM-China-Catchment-Attributes-and-Meteorology-dataset/blob/main/data/glim_cate_number_mapping.csv	./data/glim_cate_number_mapping.csv ./data/glim_name_short_long.txt	These files are used for name conversions in the program.
GDBD	https://www.cger.nies.go.jp/db/gdbd/gdbd_index_e.html	./data/river_network/as_streams_wgs.shp	River network shapefiles are used to determine river basin shape factors. The source data need to be reprojected

to EPSG:4326 (using ArcMap or QGIS) to successfully run the code. Note that files in different regions have different names.

Run the code

When all the data is ready, the user can run the code calculate_all_attributes.py to calculate all attributes or run separate scripts (e.g., soil.py) to calculate indicators for specific categories. The result will appear in the output folder.

Financial support

This research has been supported by the National Key Research and Development Program (2018YFC0407901, 2018YFC0407905), the National Natural Science Fund of China (51779100), and the Central Public-interest Scientific Institution Basal Research Fund (HKY-JBYW-2020-21, HKY-JBYW-2020-07).

References

Abrams, M., R. Crippen and H. J. R. S. Fujisada (2020). "ASTER global digital elevation model (GDEM) and ASTER global water body dataset (ASTWBD)." **12**(7): 1156.

Addor, N., H. X. Do, C. Alvarez-Garreton, G. Coxon, K. Fowler and P. A. J. H. S. J. Mendoza (2020). "Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges." **65**(5): 712-725.

Addor, N., A. J. Newman, N. Mizukami and M. P. Clark (2017). "The CAMELS data set: catchment attributes and meteorology for large-sample studies." Hydrology and Earth System Sciences (HESS) **21**(10): 5293-5313.

Alvarez-Garreton, C., P. A. Mendoza, J. P. Boisier, N. Addor, M. Galleguillos, M. Zambrano-Bigiarini, A. Lara, G. Cortes, R. Garreaud and J. McPhee (2018). "The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies-Chile dataset." Hydrology and Earth System Sciences **22**(11): 5817-5846.

Belward, A. S., J. E. Estes and K. D. Kline (1999). "The IGBP-DIS global 1-km land-cover data set DISCover: A project overview." Photogrammetric Engineering and Remote Sensing **65**(9): 1013-1020.

Berghuijs, W. R., E. E. Aalbers, J. R. Larsen, R. Trancoso and R. A. Woods (2017). "Recent changes in extreme floods across multiple continents." Environmental Research Letters **12**(11): 114035.

Blume, T., I. van Meerveld and M. Weiler (2018). "Incentives for field hydrology and data sharing: collaboration and compensation: reply to "A need for incentivizing field hydrology, especially in an era of open data"." Hydrological Sciences Journal **63**(8): 1266-1268.

Brodeur, Z. P., J. D. Herman and S. Steinschneider (2020). "Bootstrap Aggregation and Cross-Validation Methods to Reduce Overfitting in Reservoir Control Policy Search." Water Resources Research **56**(8): e2020WR027184.

Buermann, W., J. Dong, X. Zeng, R. B. Myneni and R. E. Dickinson (2001). "Evaluation of the utility of satellite-based vegetation leaf area index data for climate simulations." Journal of Climate **14**(17): 3536-3550.

Center, G. G. R. D. (2005). Global Runoff Database, Koblenz.

Ceola, S., B. Arheimer, E. Baratti, G. Blöschl, R. Capell, A. Castellarin, J. Freer, D. Han, M. Hrachowitz, Y. J. H. Hundecha and E. S. Sciences (2015). "Virtual laboratories: new opportunities for collaborative water science." **19**(4): 2101-2117.

Chagas, V. B., P. L. Chaffe, N. Addor, F. M. Fan, A. S. Fleischmann, R. C. Paiva and V. A. Siqueira (2020). "CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil." Earth System Science Data **12**(3): 2075-2096.

China, M. o. G. a. M. R. o. t. P. s. R. o. (1991). "Geological map of Nei Mongol Autonomous Region, People's Republic of China, scale 1:1,500,000."

Coron, L., V. Andreassian, C. Perrin, J. Lerat, J. Vaze, M. Bourqui and F. Hendrickx (2012). "Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments." Water Resources Research **48**(5).

Coxon, G., N. Addor, J. P. Bloomfield, J. Freer, M. Fry, J. Hannaford, N. J. Howden, R. Lane, M. Lewis and E. L. Robinson (2020). "CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain." Earth System Science Data **12**(4): 2459-2483.

Dai, Y., Q. Xin, N. Wei, Y. Zhang, W. Shangguan, H. Yuan, S. Zhang, S. Liu and X. Lu (2019). "A global high-resolution data set of soil hydraulic and thermal properties for land surface modeling." Journal of Advances in Modeling Earth Systems **11**(9): 2996-3023.

de Araújo, J. C. and J. I. González Piedra (2009). "Comparative hydrology: analysis of a semiarid and a humid tropical watershed." Hydrological Processes: An International Journal **23**(8): 1169-1178.

Desborough, C. E. (1997). "The impact of root weighting on the response of transpiration to moisture stress in land surface schemes." Monthly Weather Review **125**(8): 1920-1930.

Didan, K. (2015). MOD13A3 MODIS/Terra vegetation indices monthly L3 global 1km SIN grid V006 (Data set) NASA EOSDIS Land Process, DAAC.

Feng, D., K. Fang and C. Shen (2020). "Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales." Water Resources Research **56**(9): e2019WR026793.

Friedl, M. A., D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley and X. Huang (2010). "MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets." Remote sensing of Environment **114**(1): 168-182.

Gleeson, T., N. Moosdorf, J. Hartmann and L. Van Beek (2014). "A glimpse beneath earth's surface: GLobal HYdrogeology MaPS (GLHYMPS) of permeability and porosity." Geophysical Research Letters **41**(11): 3891-3898.

Gleeson, T., L. Smith, N. Moosdorf, J. Hartmann, H. H. Dürr, A. H. Manning, L. P. van Beek and A. M. Jellinek (2011). "Mapping permeability over the surface of the Earth." Geophysical Research Letters **38**(2).

Gudmundsson, L., M. Leonard, H. X. Do, S. Westra and S. I. Seneviratne (2019). "Observed trends in global indicators of mean and extreme streamflow." Geophysical Research Letters **46**(2): 756-766.

Hartmann, J. and N. Moosdorf (2012). "The new global lithological map database GLiM: A representation of rock properties at the Earth surface." Geochemistry, Geophysics, Geosystems **13**(12).

Hengl, T., J. Mendes de Jesus, G. B. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng and B. Bauer-Marschallinger (2017). "SoilGrids250m: Global gridded soil information based on machine learning." PLoS one **12**(2): e0169748.

Horn, B. K. (1981). "Hill shading and the reflectance map." Proceedings of the IEEE **69**(1): 14-47.

Huang, H., Y. Han, M. Cao, J. Song and H. Xiao (2016). "Spatial-temporal variation of aridity index of China during 1960–2013." Advances in Meteorology **2016**.

Jenson, S. K., J. O. J. P. e. Domingue and r. sensing (1988). "Extracting topographic structure from digital elevation data for geographic information system analysis." **54**(11): 1593-1600.

Kendall, M. G. J. B. (1938). "A new measure of rank correlation." **30**(1/2): 81-93.

Knoben, W. J., J. E. Freer and R. A. Woods (2019). "Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores." Hydrology and Earth System Sciences **23**(10): 4323-4331.

Knyazikhin, Y. (1999). "MODIS leaf area index (LAI) and fraction of photosynthetically active radiation absorbed by vegetation (FPAR) product (MOD 15) algorithm theoretical basis document." http://eospso. gsfc. nasa. gov/atbd/modistabls. heml.

Kollat, J., P. Reed and T. Wagener (2012). "When are multiobjective calibration trade-offs in hydrologic models meaningful?" Water Resources Research **48**(3).

Kollat, J., P. Reed and T. J. W. R. R. Wagener (2012). "When are multiobjective calibration trade-offs in hydrologic models meaningful?" **48**(3).

Kratzert, F., D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter and G. Nearing (2019). "Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets." Hydrology & Earth System Sciences **23**(12).

Lane, R. A., G. Coxon, J. E. Freer, T. Wagener, P. J. Johnes, J. P. Bloomfield, S. Greene, C. J. Macleod and S. M. Reaney (2019). "Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1000 catchments in Great Britain." Hydrology and Earth System Sciences **23**(10): 4011-4032.

Legasa, M. and J. M. Gutiérrez (2020). "Multisite Weather Generators using Bayesian Networks: An illustrative case study for precipitation occurrence." Water Resources Research **56**(7): e2019WR026416.

Lehner, B. (2014). HydroBASINS: Global watershed boundaries and sub-basin delineations derived from HydroSHEDS data at 15 second resolution—Technical documentation version 1. c.

Lehner, B., C. R. Liermann, C. Revenga, C. Vörösmarty, B. Fekete, P. Crouzet, P. Döll, M. Endejan, K. Frenken and J. J. T. D. Magome, Version (2011). "Global reservoir and dam (grand) database." **1**: 1-14.

Linke, S., B. Lehner, C. O. Dallaire, J. Ariwi, G. Grill, M. Anand, P. Beames, V. Burchard-Levine, S. Maxwell and H. Moidu (2019). "Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution." Scientific data **6**(1): 1-15.

Liu, B., M. Xu, M. Henderson and W. Gong (2004). "A spatial analysis of pan evaporation trends in China, 1955–2000." Journal of Geophysical Research: Atmospheres **109**(D15).

Liu, Q., Z. Yang and X. Xia (2010). "Trends for pan evaporation during 1959-2000 in China." Procedia Environmental Sciences **2**: 1934-1941.
Liu, Y., J. Zheng, Z. Hao and X. Zhang (2017). "Unprecedented warming revealed from multi-proxy reconstruction of temperature in southern China for the past 160 years." Advances in Atmospheric Sciences **34**(8): 977-982.
Maidment, D. R. (1996). GIS and hydrologic modeling-an assessment of progress. Third International Conference on GIS and Environmental Modeling, Santa Fe, New Mexico.
Maidment, D. R. and S. Morehouse (2002). Arc Hydro: GIS for water resources, ESRI, Inc.
Masutomi, Y., Y. Inui, K. Takahashi and Y. Matsuoka (2009). "Development of highly accurate global polygonal drainage basin data." Hydrological Processes: An International Journal **23**(4): 572-584.
Mei, Y., V. Maggioni, P. Houser, Y. Xue and T. Rouf (2020). "A nonparametric statistical technique for spatial downscaling of precipitation over High Mountain Asia." Water Resources Research **56**(11): e2020WR027472.
Myneni, R., Y. Knyazikhin and T. Park (2015). "MOD15A2H MODIS/terra leaf area index/FPAR 8-day L4 global 500 m SIN grid V006." NASA EOSDIS Land Processes DAAC.
Nevo, S., V. Anisimov, G. Elidan, R. El-Yaniv, P. Giencke, Y. Gigi, A. Hassidim, Z. Moshe, M. Schlesinger and G. Shalev (2019). "ML for flood forecasting at scale." arXiv preprint arXiv:1901.09583.
Newman, A., M. Clark, K. Sampson, A. Wood, L. Hay, A. Bock, R. Viger, D. Blodgett, L. Brekke and J. Arnold (2015). "Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance." Hydrology and Earth System Sciences **19**(1): 209-223.
Ni, H. and S. M. Benson (2020). "Using Unsupervised Machine Learning to Characterize Capillary Flow and Residual Trapping." Water Resources Research **56**(8): e2020WR027473.
Oudin, L., V. Andréassian, J. Lerat and C. Michel (2008). "Has land cover a significant impact on mean annual streamflow? An international assessment using 1508 catchments." Journal of hydrology **357**(3-4): 303-316.
Running, S., Q. Mu and M. Zhao (2017). MOD16A2 MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500 m SIN Grid V006. NASA EOSDIS Land Processes DAAC.
Seybold, H., D. H. Rothman and J. W. Kirchner (2017). "Climate's watermark in the geometry of stream networks." Geophysical Research Letters **44**(5): 2272-2280.
Shangguan, W., Y. Dai, Q. Duan, B. Liu and H. Yuan (2014). "A global soil data set for earth system modeling." Journal of Advances in Modeling Earth Systems **6**(1): 249-263.
Shangguan, W., Y. Dai, B. Liu, A. Zhu, Q. Duan, L. Wu, D. Ji, A. Ye, H. Yuan and Q. Zhang (2013). "A China data set of soil properties for land surface modeling." Journal of Advances in Modeling Earth Systems **5**(2): 212-224.
Shen, C., E. Laloy, A. Elshorbagy, A. Albert, J. Bales, F.-J. Chang, S. Ganguly, K.-L. Hsu, D. Kifer and Z. Fang (2018). "HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community." Hydrology and Earth System Sciences (Online) **22**(11).
Silberstein, R. (2006). "Hydrological models are so good, do we still need data?" Environmental Modelling & Software **21**(9): 1340-1352.
Singh, R., S. Archfield and T. Wagener (2014). "Identifying dominant controls on hydrologic parameter transfer from gauged to ungauged catchments–A comparative hydrology approach." Journal of Hydrology **517**: 985-996.
Singh, R., K. van Werkhoven and T. Wagener (2014). "Hydrological impacts of climate change in gauged and ungauged watersheds of the Olifants basin: a trading-space-for-time approach." Hydrological Sciences Journal **59**(1): 29-55.
Subramanya, K. (2013). Engineering Hydrology, 4e, Tata McGraw-Hill Education.
Sulla-Menashe, D. and M. A. Friedl (2018). "User guide to collection 6 MODIS land cover (MCD12Q1 and MCD12C1) product." USGS: Reston, VA, USA: 1-18.
Survey, C. G. (2001). "1:2,500,000-scale digital geological map database of China."
Tyralis, H., G. Papacharalampous and S. Tantanee (2019). "How to explain and predict the shape parameter of the generalized extreme value distribution of streamflow extremes using a big dataset." Journal of Hydrology **574**: 628-645.
van Werkhoven, K., T. Wagener, P. Reed and Y. J. W. R. R. Tang (2008). "Characterization of watershed model behavior across a hydroclimatic gradient." **44**(1).
van Wijk, M. T. and M. Williams (2005). "Optical instruments for measuring leaf area index in low vegetation: application in arctic ecosystems." Ecological Applications **15**(4): 1462-1470.
Voepel, H., B. Ruddell, R. Schumer, P. A. Troch, P. D. Brooks, A. Neal, M. Durcik and M. Sivapalan (2011). "Quantifying the role of climate and landscape characteristics on hydrologic partitioning and vegetation response." Water Resources Research **47**(10).
Wang, J., M. Chen, G. Lü, S. Yue, Y. Wen, Z. Lan and S. Zhang (2020). "A data sharing method in the open web environment: Data sharing in hydrology." Journal of Hydrology **587**: 124973.
Wickel, B., B. Lehner and N. Sindorf (2007). HydroSHEDS: A global comprehensive hydrographic dataset. AGU Fall Meeting Abstracts.
Wongso, E., R. Nateghi, B. Zaitchik, S. Quiring and R. Kumar (2020). "A Data-Driven Framework to Characterize State-Level Water Use in the United States." Water Resources Research **56**(9): e2019WR024894.
Woods, R. A. J. A. i. W. R. (2009). "Analytical model of seasonal climate impacts on snow hydrology: Continuous snowpacks." **32**(10): 1465-1481.

Xinjiang, B. o. G. a. M. R. o. (1992). "Geological map of Xinjiang Uygur, Autonomous Region, China, version 2, scale 1:1,500,000."

Xu, Y., X. Gao, Y. Shen, C. Xu, Y. Shi and a. Giorgi (2009). "A daily temperature dataset over China and its application in validating a RCM simulation." Advances in Atmospheric sciences **26**(4): 763-772.

Yamazaki, D., D. Ikeshima, J. Sosa, P. D. Bates, G. H. Allen and T. M. Pavelsky (2019). "MERIT Hydro: a high-resolution global hydrography map based on latest topography dataset." Water Resources Research **55**(6): 5053-5073.

Zeng, X. (2001). "Global vegetation root distribution for land modeling." Journal of Hydrometeorology **2**(5): 525-530.