Corrected ERA5 precipitation by machine learning significantly improved flow simulations for the Third Pole basins

He Sun[1*], Tandong Yao[1,2,3], Fengge Su[1,2,3], Zhihua He[4], Guoqiang Tang[5], Ning Li[1,3], Bowen Zheng[1,3], Jingheng Huang[1,3], Fanchong Meng[6], Tinghai Ou[7], Deliang Chen[7]

[1]State Key Laboratory of Tibetan Plateau Earth System, Resources and Environment, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China

[2]CAS Center for Excellence in Tibetan Plateau Earth Sciences, Beijing 100101, China

[3]University of Chinese Academy of Sciences, Beijing 100101, China

[4]Centre for Hydrology, University of Saskatchewan, Saskatoon, Saskatchewan S7N 1K2, Canada

[5]Centre for Hydrology, University of Saskatchewan, Canmore, Alberta T1W 3G1, Canada

[6]College of Geosciences and Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450011, China

[7]Regional Climate Group, Department of Earth Sciences, University of Gothenburg, Gothenburg 405 30, Sweden

*Corresponding author: He Sun

State Key Laboratory of Tibetan Plateau Earth System, Resources and Environment, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China

Email: sunhe@itpcas.ac.cn

Abstract

Precipitation is one of the most important atmospheric inputs to hydrological models. However, existing precipitation datasets for the Third Pole (TP) basins show large discrepancies in precipitation magnitudes and spatiotemporal patterns, which poses a great challenge to hydrological simulations in the TP basins. In this study, a gridded (10 km×10 km) daily precipitation dataset is constructed through a random forest-based machine learning algorithm (RF algorithm) correction of the ERA5 precipitation estimates based on 940 gauges in 11 upper basins of TP for 1951–2020. The data set is evaluated by gauge observations at point scale, and is inversely evaluated by the Variable Infiltration Capacity hydrological model linked with a glacier melt algorithm (VIC-Glacier). The corrected ERA5 (ERA5_cor) agrees well with gauge observations after eliminating the severe overestimation in the original ERA5 precipitation. The corrections greatly reduce the original ERA5 precipitation estimates by 10%–50% in 11 basins of the TP, and present more details on precipitation spatial variability. The inverse hydrological model evaluation demonstrates the accuracy and rationality, and we provide an updated estimate of runoff components contribution to total runoff in seven upper basins in the TP based on the VIC-Glacier model simulations with the ERA5_cor precipitation. This study provides good precipitation estimates with high spatiotemporal resolution for 11 upper basins in the TP, which are expected to facilitate the hydrological modeling and prediction studies in this high mountainous region.

Keywords

Precipitation correction; ERA5 precipitation; Hydrological model; Random Forest algorithm; Tibetan Plateau

Plain Language Summary

The Third Pole (TP) is the source of water to the people living in the downstream. Precipitation is the key driver of terrestrial hydrological cycle and the most important atmospheric input to land surface hydrological models. However, none of current precipitation data are equally good for all the TP basins because of high variabilities in their magnitudes and spatiotemporal patterns, posing a great challenge to the hydrological simulation. Therefore, in this study, a gridded daily precipitation dataset (10 km×10 km) is reconstructed through RF algorithm correction of ERA5 precipitation estimates based on 940 gauges in 11 TP basins for 1951–2020. The data eliminates the severe overestimation of original ERA5 precipitation estimates and presents more reasonable spatial variability, and also exhibits a high potential for hydrological application in the TP basins. This study provides a long-term precipitation data for climate and hydrological studies, and provides a reference for deriving precipitation in high mountainous regions with complex terrain and limited observations.

1 Introduction

Precipitation is the key forcing variable for the land surface hydrological process at a variety of space-timescales; therefore, accurate precipitation input is crucial for reliable hydrological simulations. This is particularly important in orographically influenced high-mountain terrains, which are highly sensitive and prone to climate change (Viviroli et al., 2011). Unfortunately, direct precipitation observations are often either sparse or nonexistent in high mountainous regions. The Third Pole (TP) is the high-elevation area in Asia centered on the Tibetan Plateau, which is also the origin of major Asian rivers (Fig. 1). The TP is characterized by a lack of meteorological observations and heterogeneous distribution of stations due to the complex topography and harsh environmental conditions. For instance, availably long-term and continuous national meteorological stations are mostly located in the low-altitude valleys of the eastern TP, and uncertainties exist in the estimation due to the orographic influence and wind-induced undercatch of gauge estimates (Yang et al., 2005).

Apart from direct observations, gridded precipitation products including satellite, reanalysis and regional climate models provide alternative sources of estimates in the TP. However, existing studies on multiple precipitation evaluations in TP basins suggest that none of the estimates is accurate for all basins and under all conditions, given the large differences in magnitudes and spatiotemporal characteristics. Tong et al. (2014b) compared widely used gauge-based, reanalysis and satellite-based estimates with corrected gauge observations in the TP, and showed that all the datasets could detect the large-scale precipitation regime, but they performed differently in detecting mean annual precipitation in the basins. Tang et al. (2020) compared nine satellite and reanalysis precipitation products for China and found that all products are unsatisfactory in the TP due to its more complex climate/topography conditions and lower rain gauge density compared to other regions in China. Sun et al. (2021a) evaluated the performance of precipitation estimates from the weather research and forecasting (WRF) model in the TP basins during 2001–2008, and suggested that it presented more details on the spatial pattern, but tended to overestimate the mean annual estimates from gauge-based estimates by 20–95%. Dahri et al. (2021) evaluated 27 gridded precipitation datasets in the high-altitude Indus, and indicated marked differences in spatiotemporal and quantitative distribution of precipitation among the data sets. It is worth noting that among all the products, the newly released fifth-generation reanalysis (ERA5) precipitation of the European Centre for Medium-Range Weather Forecasts (ECMWF) performed the best among these datasets in the upper Indus (Dahri et al., 2021; Lai et al., 2021). However, a recent evaluation of ERA5 in 11 basins in the TP (Sun et al., 2021b) suggested that ERA5 precipitation tended to overestimate the gauge observations by 30%–270%, and thus general overestimations (33%–106%) are existed in all basins except for the upper Indus when used to simulate streamflow by a hydrological model.

Glacier and snow melts can significantly modify streamflow regimes in high mountainous basins (Bookhagen and Burbank, 2010). The uncertainties in current precipitation datasets largely influence the quantitative assessment of the impacts of glacier and snow meltwater on streamflow. For instance, Khanal et al. (2021) simulated streamflow directly driven by the ERA5 precipitation without considering its uncertainty, and estimated glacier runoff contributed about 5.1% in the high-altitude Indus, which was 40.6% lower than a previous estimates by the same hydrological model with different precipitation input (Lutz et al., 2014). Another example is the Yarlung Zangbo basin, which has the largest glacier area percentage (1.5%, Table 1) among the monsoon-dominated basins. Zhao et al. (2019) estimated glacier runoff contributions of 5.5% to the total runoff in the Nuxia hydrological station of the Yarlung Zangbo river basin, while Sun and Su (2020) concluded that glacier runoff simulated with a reconstructed precipitation dataset contributed 14%–16% to the total runoff in this basin.

Given the large uncertainties of widely used precipitation estimates in hydrological simulation in the TP, bias corrections for gridded precipitation estimates are needed before they are used as inputs for hydrological modeling. However, suitable correction approaches and the density of rain gauges are two major limitations for precipitation corrections in the TP. Recently, machine learning algorithms based on statistical learning theory have been applied to simulate or downscale key variables involved in the hydrometeorological cycle (He et al., 2021; Jiang et al., 2021; Oppel and Fischer, 2020; Wang et al., 2020; Zhang and Ye, 2021). These advantages of adequate interactions among different variables, small requirements of parameter-tuning and a low chance of overfitting in machine learning algorithms provide an opportunity to correct precipitation data in the TP. In addition, the reliability of corrected precipitation estimates by different correction approaches is highly dependent on the density of observed stations. Aiming to address this issue, we have tried our best to collect precipitation gauge observations from different projects (i.e., the Second Tibetan Plateau Scientific Expedition and Research project), multiple research institutions, and relevant governmental hydrometeorological agencies. These gauge data constitute a unique observation basis for precipitation correction.

In this study, we intend to correct the daily ERA5 precipitation for 1951–2020 based on the gauge observations by using machine learning correction algorithm for 11 upper basins in the TP (Fig. 1), including Yangtze (UYA), Yellow (UYE), Lancang (ULC), Nujiang (UNJ) and Yarlung Zangbo (YZ) in the monsoon-dominated TP regions, and Indus (UI), Amu Darya (UAMD), Syr Darya (USRD), Yarkant (UYK), Hotan (UHT) and Aksu (UAKS) in the westerlies-dominated TP regions. The corrected precipitation is evaluated at point scales, and is inversely evaluated by the Variable Infiltration Capacity (VIC) land surface hydrological model linked with a temperature-index model (VIC-Glacier) by comparison with observed streamflow. This study aims to provide accurate precipitation estimates with high spatiotemporal resolution for hydrological simulations in major TP basins.

2 Study area

The TP is the high-elevation area in Asia centered on the Tibetan Plateau and surroundings, with a total area of~5,000,000 km^2 and mean elevation of 4000 m. It extends from the Himalaya in the south to Kunlun and Qilian mountains in the north, and presents a west-east span from Pamirs mountains and the Hindu Kush in the west to Hengduan mountains in the east (Yao, 2014). The study basins of UYA, UYE, ULC, UNJ, YZ, UI, UYK, UAKS, UHT, UAMD and USRD here are defined as all regions upstream of the hydrological stations, respectively (Fig. 1; Table 1). The UAKS basin is comprised of two branches controlled by the Shaliguilanke and Xiehela hydrological stations, respectively. And the UHT basin is controlled by the Wuluwati and Toktogul stations, respectively. The UYE, UYA, ULC, UNJ and YZ basins are located in the monsoon-dominated southeastern TP, with more than 70% of annual precipitation occurring in June–September (Sun et al., 2021a). The remaining six basins (the UI, UYK, UAKS, UHT, UAMD, and USRD) are mostly affected by the westerlies system. The UI, UAMD, and USRD have similar precipitation regimes with more than 70% of annual precipitation occurring in November–April, and more than 55% of annual precipitation occurs in May–August over the UYK, UAKS and UHT basin (Kan et al., 2018; Mölg et al., 2013). The glacier distributions are highly uneven among the basins with a larger coverage in the westerlies-dominated basins (2.5%–20%) than in the monsoon-dominated basins (below 1.5%).

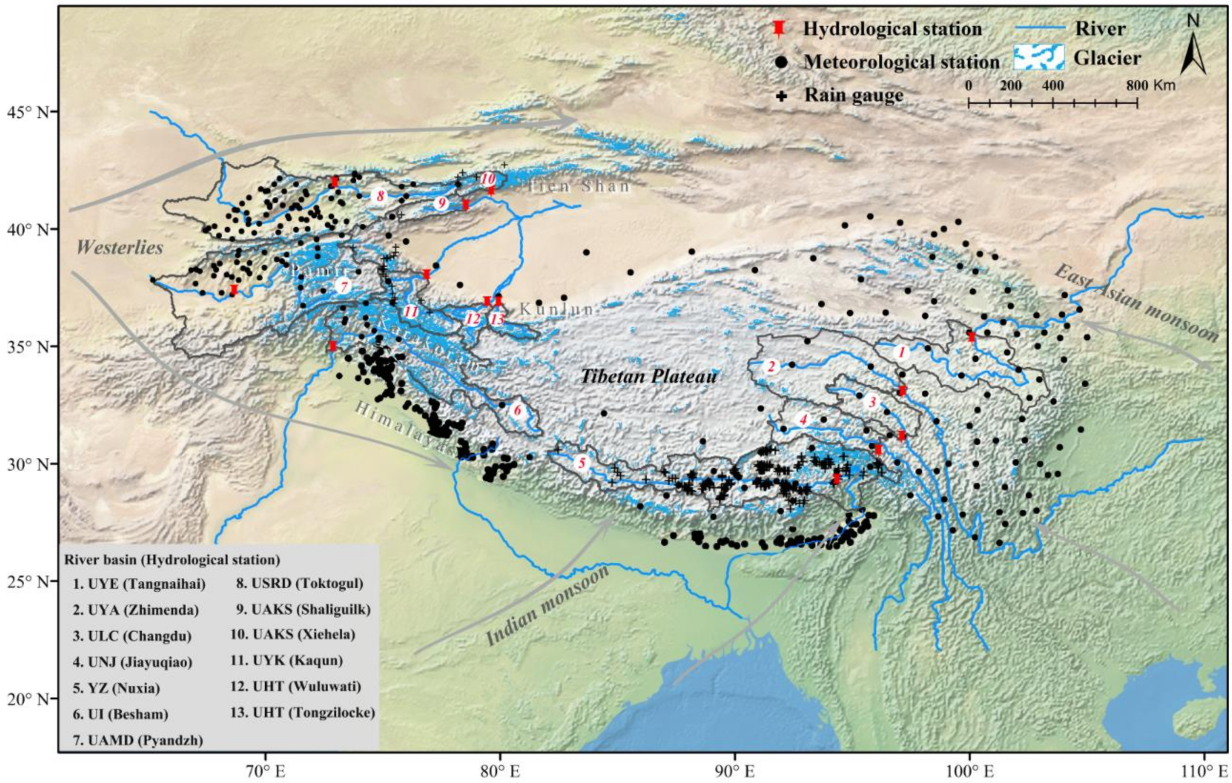

Fig. 1. Topography and boundaries of 11 upper river basins in the Third Pole. The sequence numbers 1–13 denote the hydrological stations of the upper regions of the Yellow (UYE), Yangtze (UYA), Lancang (ULC), Nujiang (UNJ), Yarlung Zangbo (YZ), Indus (UI), Amu Darya (UAMD), Syr Darya (USRD), two branches of the Aksu (UAKS), Yarkant (UYK) and two branches of the Hotan (UHT) river basins, respectively.Meteorological stations and rain gauges are represented with black points and crosses, respectively. The red pushpins denote the hydrological stations used in this study. The base map of topography is from the Natural Earth (https://www.naturalearthdata.com/).

Table 1. Characteristics of 11 upstream river basins in the Third Pole. UYE, UYA, ULC, UNJ, YZ, UI, UAMD, USRD, UYK, UAKS and UHT represent the upstream regions of each control station of the Yellow, Yangtze, Lancang, Nujiang, Yarlung Zangbo, Indus, Amu Darya, Syr Darya, Yarkant, Aksu and Hotan river basins, respectively (Also see Figure 1).

	Basin	Hydrological Station		Latitude	Longitude	Glacier coverage
		Name	No.	(°N)	(°E)	(%)
Monsoon-dominated	UYA	Zhimenda	1	33.02	97.13	0.8
	UYE	Tangnaihai	2	35.30	100.09	0.1
	ULC	Changdu	3	31.11	97.11	0.3
	UNJ	Jiayuqiao	4	30.51	96.12	1.2
	YZ	Nuxia	5	29.47	94.57	1.5
Westerlies-dominated	UI	Besham	6	34.92	72.88	11.9
	USRD	Toktogul	7	41.90	72.95	2.5
	UAMD	Pyandzh	8	37.33	68.67	5.3
	UYK	Kaqun	11	37.98	76.90	9.3
	UAKS	Shaliguilk	9	78.54	40.94	3.5
		Xiehela	10	79.62	41.57	18.6
	UHT	Wuluwati	12	79.44	36.84	9.2
		Toktogul	13	79.92	36.82	19.6

3 Data and method

3.1 Data

Gauge precipitation observation For the monsoon-dominated eastern and southeastern TP regions, long-term daily observations from 150 meteorological stations inside China for 1961–2016 are collected from the China Meteorological Administration (CMA, http://data.cma.cn/), and extra monthly observations from 118 meteorological stations outside China are collected from the Global Historical Climatology Network (GHCN, https://www.ncdc.noaa.gov/ghcn-monthly) for 2005–2013. In addition, monthly precipitation data for 2014–2016 from 312 rain gauges in the southeastern TP are collected from the governmental hydrometeorological agencies, which had been used in precipitation correction in the Yarlung Zangbo river basin (Sun and Su, 2020).

For the westerlies-dominated regions, long-term daily observations from 17 meteorological stations for 1961–2016 are collected from the CMA. Monthly observations from 316 meteorological stations for 1961–2000 are collected from the Global Historical Climatology Network (GHCN, https://www.ncdc.noaa.gov/ghcn-monthly). A field research campaign during 2014–2017 that installed 27 rain gauges at different altitudes for the upper Yarkant basin Monthly provided monthly observations (Kan et al., 2018).

These obtained gauge estimates have undergone quality control procedures to pre-process (either validated, corrected or removed) erroneous data (e.g., daily precipitation values less than 0 mm), and only monthly records that are derived from at least three-year's of consecutive observation are used in this study. Therefore, these data are directly used in this study without further bias correction.

Gridded precipitation estimates Daily gridded precipitation estimates from the ERA5 and WRF are used to generate the precipitation background field. The ERA5 provides hourly precipitation estimates from 1950 to the present and spatial resolution of about 30 km (Hersbach et al., 2020). It uses one of the most recent versions of the Earth system model and data assimilation method applied at ECMWF, which enables it to use modern parameterizations of Earth processes. The ERA5 precipitation data are available from https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5. The ERA5 precipitation estimates generally captures the seasonal and broad spatial distributions of gauge precipitation in TP (Sun et al., 2021b). However, ERA5 tends to overlook detailed information due to its relatively coarse resolution. Daily precipitation estimates from the WRF model, developed by the Regional Climate Group at the University of Gothenburg, provide detailed, process-based precipitation fields with a 9 km resolution (the WRF-9km hereafter) for the entire TP for 1980–2014 (Ou et al., 2020, http://biggeo.gvc.gu.se/TPReanalysis/). The WRF-9km is simulated using a continuous integration forcing strategy driven by ERA5 estimates, with spectral nudging to prevent the simulation from drifting away from the large-scale driving fields. Relative to the gauge-based estimates, the WRF-9km precipitation estimates show more detailed information on the spatial pattern (Sun et al., 2021a), but it shows large bias in the precipitation seasonality in comparison to observations in the westerlies-dominated Tarim basins (Sun et al., 2020; Sun et al., 2021). In summary, high spatial-resolution WRF-9km for 1980–2014 and high temporal-resolution ERA5 precipitation estimates for 1951–2020 are used in this study. To facilitate direct correction between WRF-9km and ERA5 precipitation datasets and keep consistent with our hydrological model setup, they are regridded to 10×10 km grids using the nearest neighbour method.

Other gridded estimates Topography, convective available potential energy, lifting condensation level and total column water vapor data in the TP basins are used in this study as inputs for the random forest-based machine learning algorithm (RF algorithm). The topography data is obtained from the United States Geological Survey data center, which is available from the CGIAR-CSI SRTM 90m database (http://srtm.csi.cgiar.org). The monthly convective available potential energy, lifting condensation level, and total column water vapor data with a 30 km resolution for 1951–2020 in the TP basins are obtained from the ERA5 datasets. These data sets are also regridded to 10×10 km grids using the nearest neighbour method to keep consistent with gridded precipitation estimates.

3.2 Precipitation correction method

The correction approach consists of three main steps (Fig. 2): first, the gridded ERA5 precipitation estimates for 1951–2020 is downscaled by the WRF estimates to generate a finer spatial pattern with a spatial resolution of 10×10 km to present more detailed information in the 11 TP basins; then, the gauge-based precipitation background field in 11 TP basins is generated by the RF algorithm based on mean monthly precipitation estimates from 580 stations in five monsoon-dominated TP regions and 360 stations in six westerlies-dominated TP regions. Finally, the downscaled ERA5 precipitation estimates on each of the grids during 1951–2020 are corrected by mean monthly gauge-based precipitation estimates. Flowchart of the correction procedure is presented in Fig. 2.

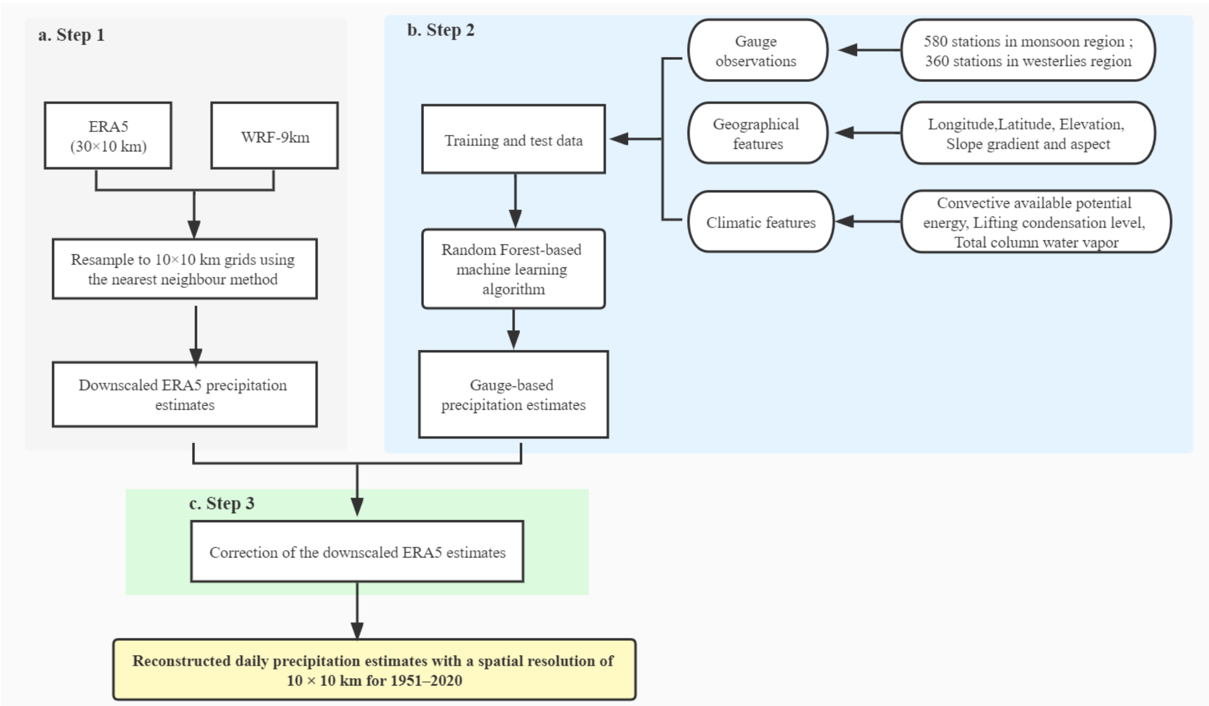

Fig. 2. Flowchart of reconstructed precipitation estimates for 1951–2020 in 11 TP basins.

3.2.1 The downscaling approach for the original ERA5

The spatial pattern of ERA5 precipitation is downscaled and corrected by high-resolution WRF-9km precipitation estimates because the WRF-9km estimates usually give more realistic precipitation estimates and detailed regional information in areas with complex terrain than coarse-resolution reanalysis data, especially for estimate of annual mean precipitation (Sun et al., 2021a). The spatial correction of ERA5 precipitation is firstly implemented by Eq. 1, in which annual precipitation estimates in the WRF-9km for 1980–2014 are used to downscale the daily ERA5 precipitation to a resolution of 10×10km.

$$P_{down,d,i} = P_{ERA5,\,d,i} \times \frac{P_{WRF,\,y,i}}{P_{ERA5,\,y,i}} \quad (\mathrm{i} = 1,2, \ldots, \mathrm{n}) \tag{1}$$

where $P_{down,d,i}$ is the downscaled daily ERA5 precipitation estimate for 1951–2020 at grid *i*, $P_{ERA5,\,d,i}$ is the original daily ERA5 precipitation estimate for 1951–2020 at grid *i*, $P_{WRF,\,y,i}$ is the annual mean WRF-9km precipitation estimate for 1980–2014 at grid *i*, and $P_{ERA5,\,y,i}$ is the original annual mean ERA5 precipitation estimate for 1980–2014 at grid *i*.

3.2.2 Generation of the gauge-based precipitation background field by the RF algorithm

The RF, one of the most popular machine learning algorithms, using a combination of numerous decorrelated decision trees to make decisions or serve as a regression analysis tool, has been adopted to generate the ground gauged precipitation background field in this step. In the RF algorithm, each of the decision trees consists of nodes and leaves that are split by randomly selected subsets of the influencing features from a randomly sampled training sub-dataset (Adhikari et al., 2020; Ehsani et al., 2020; He et al., 2021). At each node, the algorithm divides the input dataset into two classes using an optimal influencing feature and its threshold. Decision trees are grown by repeating the node splitting procedure until the end when the specified maximum tree depth is reached or all of the input cases have been divided into individual classes. The RF algorithm is used in this study because of its several advantages including being capable of treating the complex nonlinear influences and strong interactions among the selected features, as well as the low chance of overfitting benefiting from its small parameter space for tuning. The RF algorithm also runs very efficiently on large data bases with comparable performance to the current machine learning algorithms, and has been chosen for climate and hydrological prediction studies in high mountainous regions where limited gauge observations are available (Oppel and Fischer, 2020; Wang et al., 2020; Zhang and Ye, 2021).

In this study, the gauge-based precipitation background field of 11 river basins is generated by the RF algorithm based on mean monthly precipitation estimates from 580 stations in monsoon-dominated TP regions and 360 stations in westerlies-dominated TP regions. Given the large area and complex atmospheric circulation systems, the mean monthly precipitation background field is generated in five monsoon-dominated basins and six westerlies-dominated basins by the RF algorithm, respectively. Inputs of the RF algorithm selected in this study include: 1) geographical features (e.g., the longitude, latitude, elevation, slope gradient and aspect), which have influences on precipitation distribution, and 2) climatic features derived from the ERA5 (e.g., the convective available potential energy, lifting condensation level, and total column water vapor), which represent the need for the generation and development of precipitation (Cuo and Zhang, 2017; Sun et al., 2020). The forcing data on gauged grids are split into a training set (70%) and a test set (30%). Location of rain gauges used as a training and testing set are shown in Fig. S1 (seen in Supporting Information). Regression relationships between the mean gauged monthly precipitation and corresponding gridded geographical and climatic features are trained and tested by the RF algorithm on the gauged grids. The trained RF algorithm is then used to estimate mean monthly precipitation on all non-gauged grids at a resolution of 10×10 km based on the spatial inputs of geographical and climatic features, resulting in precipitation background fields in all of the 11 TP basins. Relative contributions of the multiple features to the monthly precipitation are ranked by the importance of the outputs from the RF algorithm. The estimation indicates that factors of elevation, convective available potential energy (CAPE), lifting condensation level (LCL), and total column water vapor (TCWV) together contribute more than 70% of the precipitation prediction by the RF model.

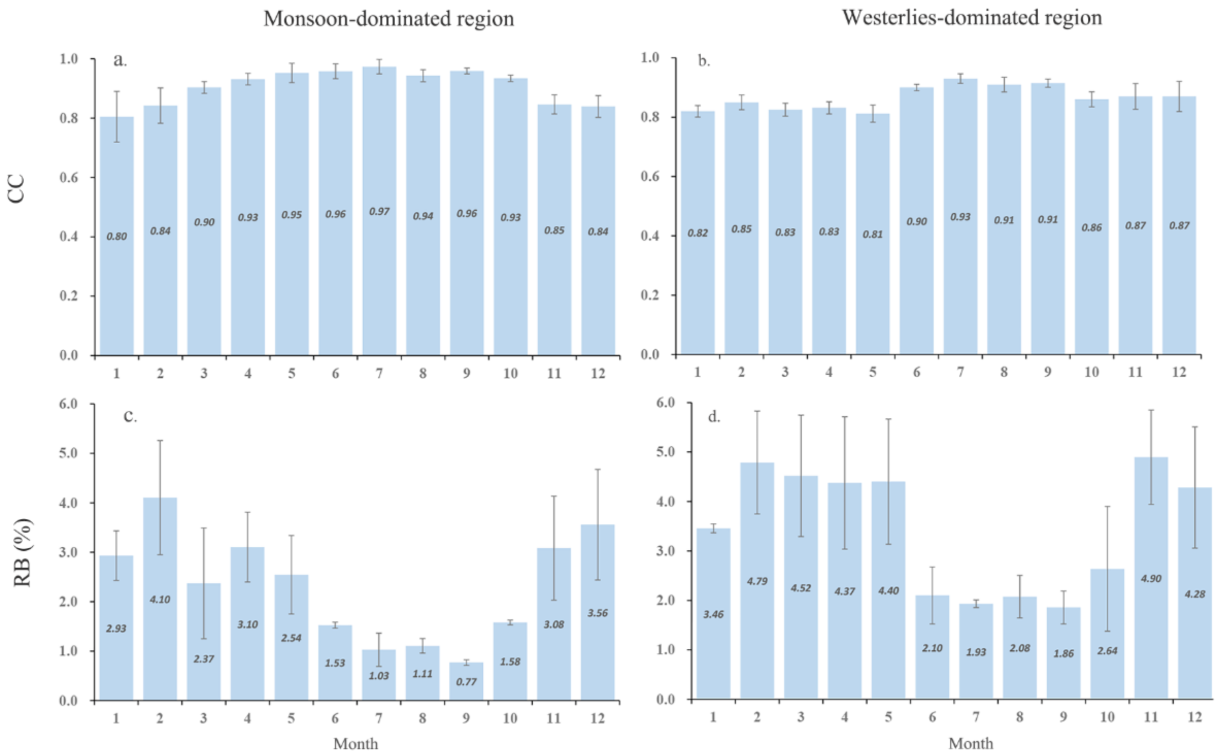

Fig. 3. Comparison between the simulations by the RF algorithm and gauge observations of the test set for each month in the monsoon- and westerlies-dominated regions of the TP, respectively.

The *ntree*, *mtry*, *node_size*, and *sample_fraction* are major tuning hyperparameters of the RF algorithm. The *ntree* is the number of decision trees, each of which is independently produced without any pruning. The *mtry* means the number of possible directions for splitting at each node of each tree. The *node_size* refers to the number of examples in each cell below which the cell is not split, and the *sample_fraction* is the number of sizes to be drawn. For the purpose of parameters tuning, the training dataset (i.e., 70% of the total gauged data) is separated into an ‘in-the-bag’ subset and an ‘out-the-bag’ subset by bootstrap samplings in the RF algorithm. The “in-the-bag” set is used for hyperparameter training. When a hyperparameter set is chosen, the RF algorithm estimates predictions for the ‘out-the-bag’ set and calculates the minimized out-of-bag mean squared error (OOB MSE). This parameter tuning procedure is repeated based on a grid search approach, and a final optimal hyperparameter set is identified when OOB MSE reaches its minimum value and stayed in stable (Janitza and Hornung 2018). In order to reduce the influence of the random sampling procedure, the RF algorithm has been repeated for 100 times for parameters tuning and testing. The *ntree* is set to 500 which ensures the OOB MSE to reach a stable level. Optimized values of *mtry*, *node size* and *sample fraction* in this study are shown in Table 2. The optimal hyperparameter set was then used in the RF algorithm to estimate predictions in the test dataset (i.e., the remaining 30% of all gauged data). Performance of the RF algorithm is assessed by comparising its precipitation prediction against observed precipitation in the test dataset using goodness of fit metrics, including the correlation coefficient (CC), root mean square error (RMSE) and relative bias (RB, %, Table 3). Relative to gauge observations of the test set, simulations by the RF algorithm show a high CC of 0.8–0.97 and a low RB of below 5% in monsoon-and westerlies-dominated basins (Fig. 3), suggesting the good reliability of the RF algorithm in estimating spatial precipitation in the basins of the TP.

Table 2. Summary of main parameters for the RF algorithm in each month in both the monsoon- and westerlies-dominated basins in the Third Pole.

Basin	Month	mtry	node size	sample fraction
Monsoon-dominated basins	Jan	3	1	1
	Feb	7	20	1
	Mar	3	13	1
	Apr	3	6	1
	May	2	1	1
	Jun	8	4	1
	Jul	4	1	1
	Aug	3	3	0.9
	Sep	3	1	0.8
	Oct	3	9	1
	Nov	4	10	0.8
	Dec	5	20	0.9
	Jan	4	3	0.9
	Feb	9	4	0.9
	Mar	4	2	1
	Apr	2	12	1

Westerlies-dominated basins	May	5	7	0.9
	Jun	3	16	0.8
	Jul	3	23	1
	Aug	3	15	0.9
	Sep	6	9	0.8
	Oct	6	2	1
	Nov	5	5	0.8
	Dec	5	3	0.8

Table 3. Statistical evaluation metrics. For the equations in this table, n is the total number of dates; and i is the ith date. O_i is the observation; S_i is the simulated streamflow; and $\overline{O}$ is the average of the observation. P_i is the precipitation estimates; and $\overline{P}$ is the average of the precipitation estimates.

Statistical Metrics	Equation
Nash-Sutcliffe efficiency coefficient (NSE)	$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(O_i - S_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O}_i)^2}$
Relative bias (RB)	$\mathrm{RB} = \frac{\sum_{i=1}^{n}(P_i - O_i)}{\sum_{i=1}^{n} O_i} \times 100\%$
Correlation coefficient (CC)	$CC = \frac{\sum_{i=1}^{n}(O_i - \overline{O})(P_i - \overline{P})}{\sqrt{\sum_{i=1}^{n}(O_i - \overline{O})^2} \times \sqrt{\sum_{i=1}^{n}(P_i - \overline{P})^2}}$
Root mean square error (RMSE)	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2}$

3.2.3 Correction of the downscaled ERA5 estimates

The downscaled ERA5 precipitation estimates on each of the 10×10 km grids during 1951–2020 are corrected by mean monthly precipitation estimated by the RF algorithm on the corresponding grid (Eq. 2). The monthly precipitation is used in this step because long-term daily precipitation estimates in most of rain gauges on the TP are very often missing. The RF algorithm is only trained by monthly precipitation observations.

$$P_{corr,d,i} = P_{down,\,d,i} \times \frac{P_{RF,\,m,i}}{P_{down,\,m,i}} \;(i = 1,2,\dots,n) \tag{2}$$

where $P_{corr,d,i}$ is the corrected ERA5 precipitation estimate for 1951–2020 at grid *i*, $P_{down,\,d,i}$ is the downscaled daily ERA5 precipitation estimate for 1951–2020 at grid *i*, $P_{RF,\,m,i}$ is the mean monthly precipitation estimate corrected by the RF algorithm at grid *i*, and $P_{down,\,m,i}$ is the downscaled mean monthly ERA5 precipitation estimate at grid *i*.

Given the correction by a systematic three-step approach, a gridded daily precipitation dataset with a spatial resolution of 10×10 km for 1951–2020 is constructed in the 11 TP basins.

3.3 Hydrological model

The Variable Infiltration Capacity (VIC) hydrological model (Liang et al., 1994; Liang et al., 1996) is a large-scale and semi-distributed macroscale land surface hydrological model that parameterizes the dominant hydrometeorological processes taking place at the land surface-atmosphere interface. In this study, the VIC linked with a simple degree-day glacier melt algorithm (Hock, 2003) termed as the VIC-Glacier model, which has been previously used in flow simulations for major river basins in the TP (Kan et al., 2018; Su et al., 2016; Sun et al., 2021a; Tong et al., 2016; Zhang et al., 2013; Zhao et al., 2019) is used to evaluate the accuracy of the corrected precipitation in 11 TP basins. The required VIC-Glacier forcing data include daily precipitation, maximum and minimum temperature, and wind speed.

In this study, the modeling frameworks at 10 km ×10 km spatial resolution, parameters and required forcing data are adopted from Zhang et al. (2013) and Sun and Su (2020) in the monsoon-dominated UYA, UYE, UNJ, ULC, and YZ basins, and from Li (2019), Kan et al.

(2018) and Huang and Su (2019) in the westerlies-dominated basins without further calibration. Therefore, the differences in simulated streamflow obtained for the same set of VIC-Glacier model parameters are entirely attributed to the differences in the precipitation inputs. In addition, to adjust the model internal stores of energy and water from the initial condition to an equilibrium state, the VIC-Glacier model is run for the years of 1951–1960 for warming up, and the years of 1961–2020 for simulation. In this study, glacier runoff is defined as all runoff generated (e.g. glacier melt and precipitation-induced runoff) in the glacierized area.

Available monthly streamflow observations from seven hydrological stations (Fig. 1, Table 1) for 1980–2010 are used to compare with simulations forced by corrected ERA5 precipitation estimates in the UYE, UYA, ULC, UNJ, YZ, UI and UYK basins, while the UAMD, USRD, UHT and UAKS basins are excluded from this comparison because of a lack of observed streamflow data for the selected time periods. The RB, CC and Nash–Sutcliffe efficiency (NSE) are used to quantify the performances between observed streamflow and simulations driven by corrected ERA5 precipitation.

4 Results

4.1 Comparison and evaluation of the corrected precipitation

Fig. 4 shows mean annual precipitation estimates from gauge observations and corresponding corrected ERA5 (ERA5_cor), downscaled ERA5 (ERA5_down) and original ERA5 grids at point scales in both monsoon- and westerlies-dominated basins in the TP, respectively. The ERA5_cor precipitation estimates exhibits good correspondence and low bias with the gauge observations in monsoon- and westerlies-dominated basins in the TP, with the CC of 0.7–0.8 and RB of 5%–9% (Fig. 4e, f). The overestimation of ERA5_cor precipitation estimates at point scales may result from precipitation undercatch in gauge observations and the scale mismatch between gauge and gridded estimates. Relative to the original ERA5 (Fig. 4a, b), the ERA5_down has larger CC and eliminates the overestimation from 41–98% to 19–46% (from 472–622 mm to 412–516 mm in RMSE) in the original ERA5 precipitation (Fig. 4c, d). Compared with the two estimates, the severe overestimation is further eliminated in the

ERA5_cor precipitation estimates, which has lower RB and RMSE, and larger CC, suggesting the large improvement of the ERA5_cor precipitation estimates. The improvements in the ERA5_cor are occurred in the representation of spatial distribution of precipitation over the complex terrain. The ERA5 precipitation generally overestimates the gauge observations in annual means with RBs of 30%–270%. However, compared to original ERA5, the ERA5_cor performs better with the RBs of mostly within ± 35% (Figure S2).

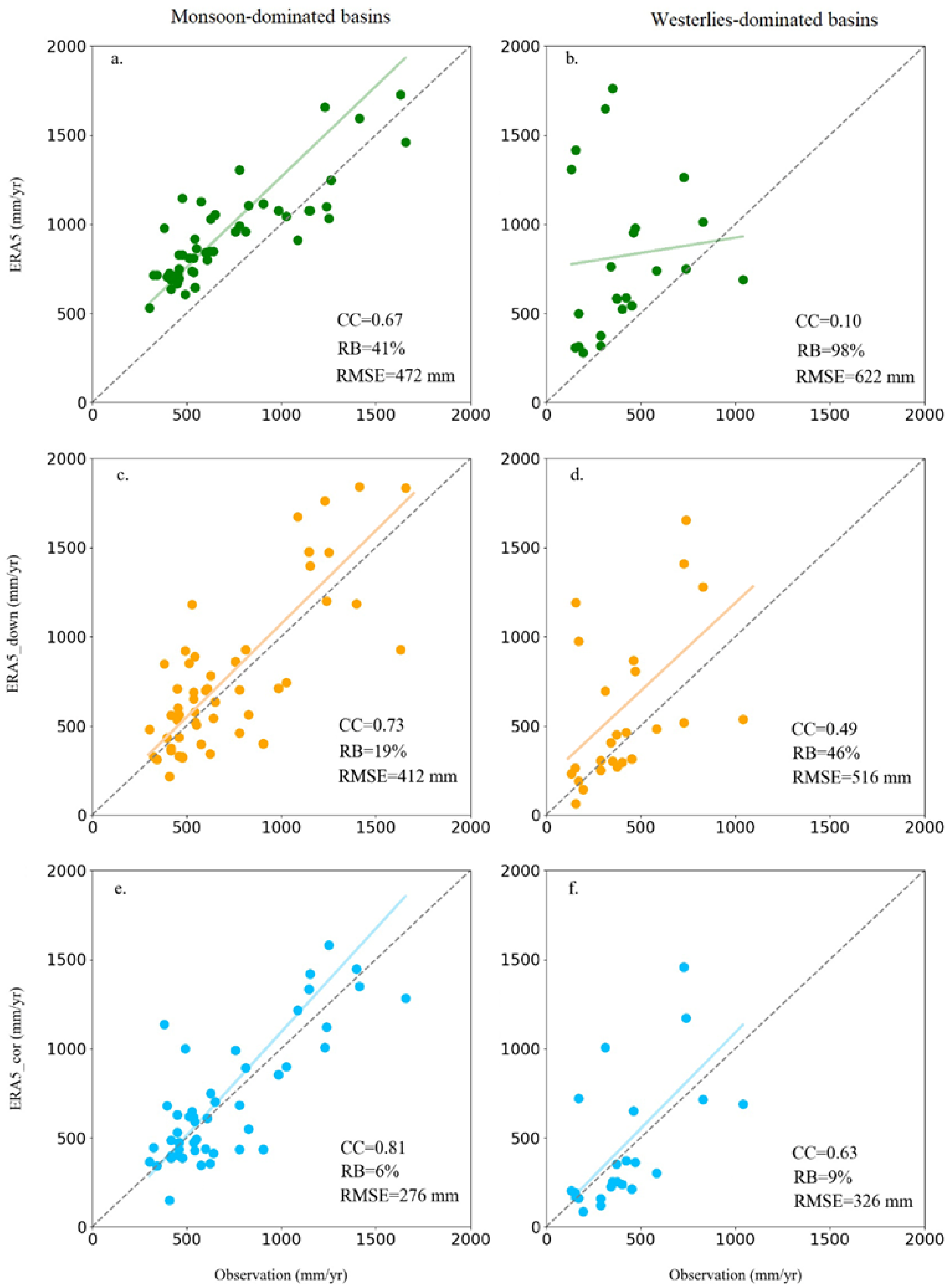

Fig. 4. Mean annual precipitation estimates from gauge observations in the test set compared to the corresponding ERA5, ERA5_down and ERA5_cor grids in the monsoon- and westerlies-dominated basins, respectively.

The large changes in the ERA5_cor also occurred in the basin-mean precipitation estimates in the TP (Fig.5). Precipitation estimates from gauge observations show consistent seasonal patterns among the monsoon basins, with more than 70% of mean annual estimates occurring in June–September (Fig. 5a–e). The westerlies-dominated UI exhibits a bimodal pattern (Fig. 5f), and the UAMD and USRD show winter-spring precipitation maximum pattern (Fig. 5g, h). It is worth noting that the UYK, UKS and UHT basin show a summer precipitation maximum. The ERA5_cor precipitation estimates successfully reproduce the seasonal pattern of gauge observations in all selected basins with CCs of above 0.9 (p<0.05) in both of the monsoon- and westerlies-dominated basins.

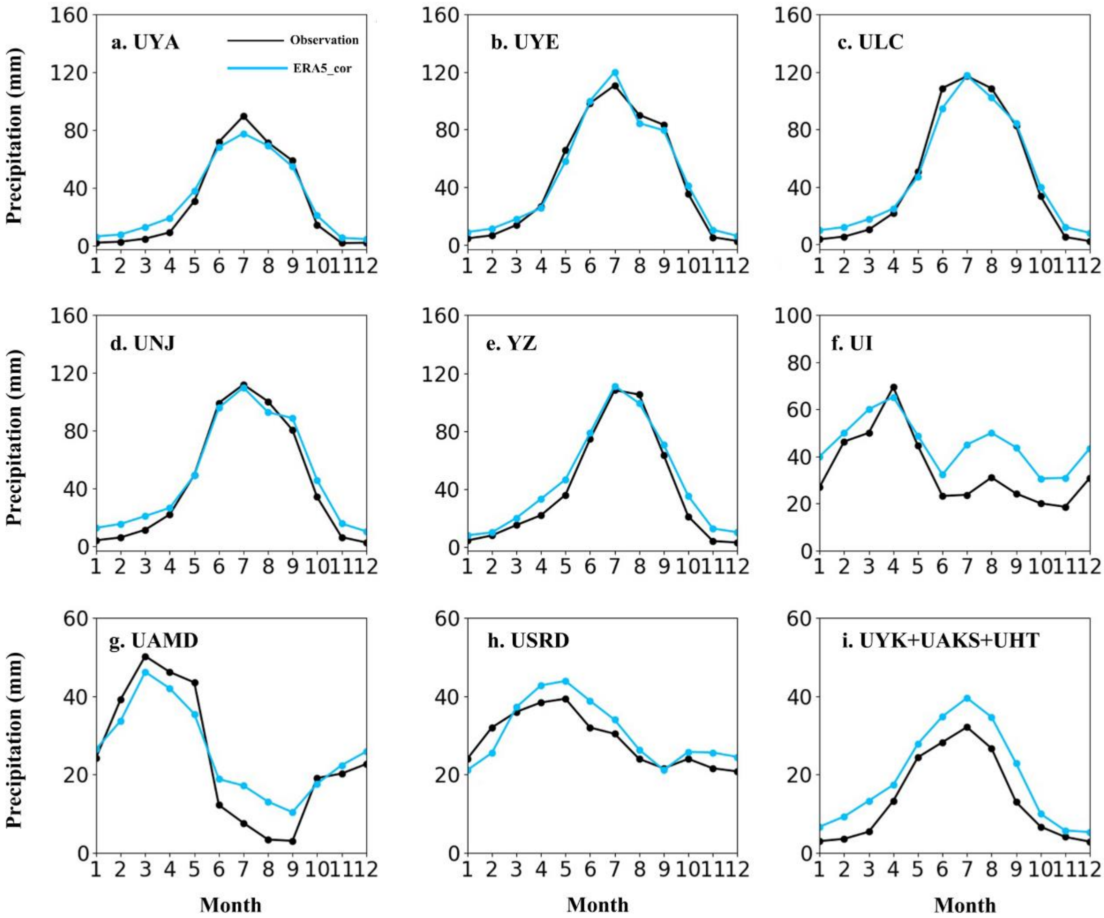

Fig. 5. Seasonal cycles of gauge observations and the corresponding ERA5_cor precipitation estimates for selected TP basins.

Fig. 6 shows seasonal cycle of the ERA5_cor and original ERA5 precipitation in 11 basins of the TP for 1951–2020. The ERA5_cor shows similar precipitation regimes to ERA5 in all basins, except for the UYK basin (Fig. 6i). Although more than 55% of the annual precipitation in both ERA5_cor and ERA5 occurs in June–August over the UYK basin, ERA5_cor precipitation peaks occur in June–July, and it is similar to the seasonal cycles of Kan et al. (2018), but ERA5 precipitation shows a peak in August.

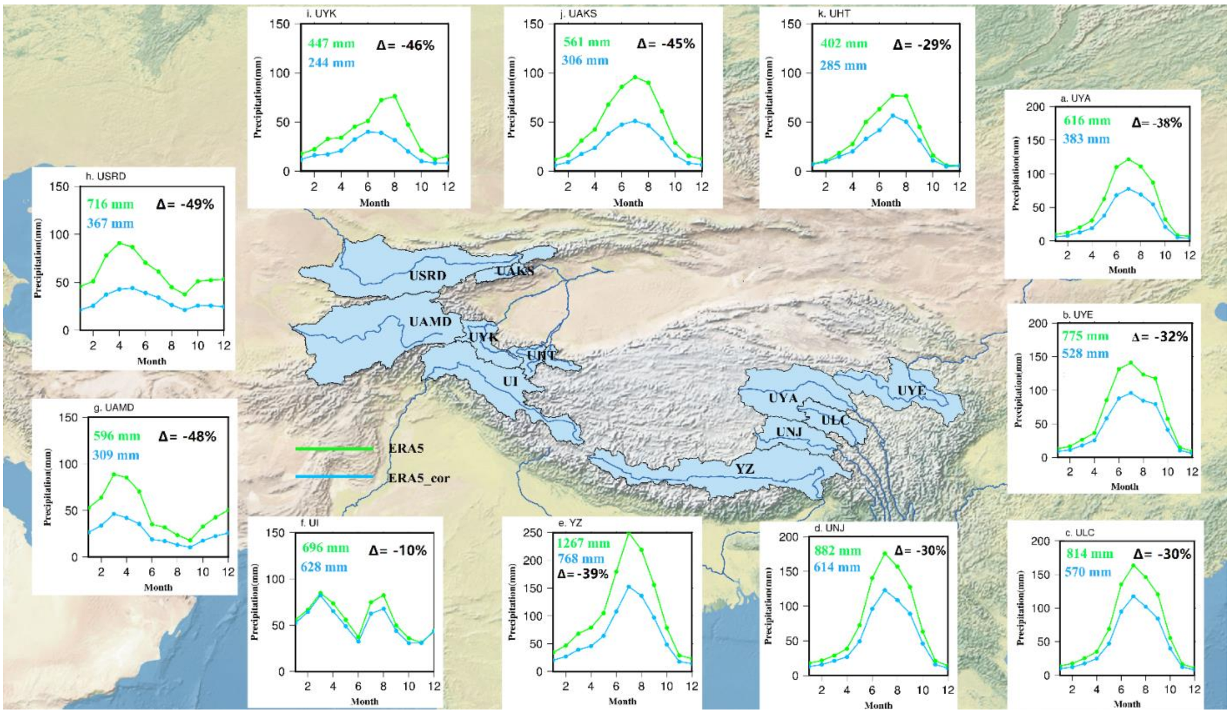

Fig. 6. Seasonal cycles of the ERA5 and ERA5_cor precipitation estimates for the 11 upper basins in the TP for 1951–2020. The numbers in mm in each panel are the mean annual precipitation estimates from ERA5 (green) and ERA5_cor (blue), respectively, and the delta represents precipitation change (%) in the ERA5_cor estimates relative to the original ERA5 estimates.

A general overestimation exists in the ERA5 precipitation in the TP basins (Sun et al., 2021b), but the ERA5_cor eliminates severe overestimations by 10–50% in the original ERA5 precipitation. The mean annual precipitation decreases from 402–1267 mm in the original ERA5 estimates to 244–768 mm in the ERA5_cor estimates. For the purpose of comparison, two precipitation estimates in the YZ and UYK basins are used as examples. Sun and Su (2020) reconstructed a precipitation dataset for 1961–2016 through precipitation gradient and linear correction methods based on 262 gauges in the YZ basin, and estimated mean annual precipitation of 709 mm. The ERA5_cor precipitation is 760 mm, which reduces the overestimation from 74% in the original ERA5 (1266 mm) to 6% after correction. Kan et al. (2018) constructed a gridded precipitation by precipitation gradient method for 1960–2015 in the UYK basin, and estimated mean annual precipitation of about 232 mm. The mean annual ERA5_cor precipitation (241 mm) is closer with this gauge-based estimates than the original ERA5 precipitation (447 mm). Overall, compared with the ERA5 precipitation estimates, lower biases in ERA5_cor suggest its validity in TP basins.

Fig. 7 shows the spatial fields of the mean annual and seasonal precipitation estimate from the ERA5 and ERA5_cor, and the difference (ERA5_cor – ERA5) between these two estimates in the 11 basins for 1961–2020. Both the ERA5 and ERA5_cor precipitation estimates can capture the large-scale spatial pattern of mean annual precipitation in the TP, with more precipitation in the monsoon-dominated basins and the highest precipitation in the southeastern TP. Both also detect the monsoon (Fig. 7b, e) and the westerlies signals (Fig. 7c, f) in TP as well.

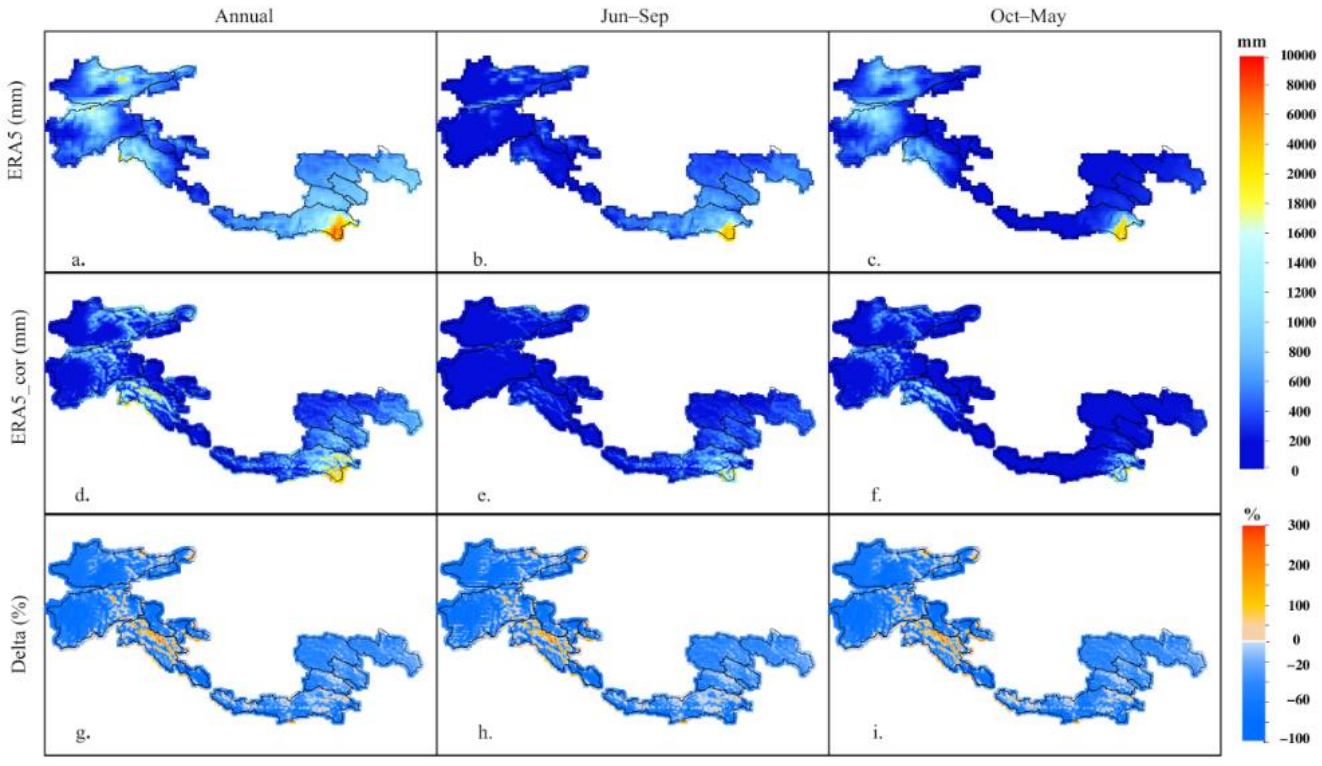

Fig. 7. Spatial fields of the mean annual and seasonal precipitation estimate from the ERA5 and ERA5_cor, and the difference (ERA5_cor -ERA5) between these two estimates in 11 basins in the TP for 1951–2020.

However, the ERA5_cor precipitation estimates present more detailed regional information on the spatial variability than that in the original ERA5 estimates due to the WRF-

9km downscaling, especially in the Yarlung Zangbo and upper Indus basins. For instance, the highest precipitation in ERA5 mainly occurs at the southern edge of the Himalayas and downstream of the UI, while the ERA5_cor precipitation shows a good resemblance to the spatial pattern of glaciers in the upper Indus. In addition, the corrected precipitation greatly increases the estimates by about 50–300% (Fig. 7g) in the high-altitude glacier area (Fig. 1), but decrease the precipitation estimates by about -30% to -90% (Fig. 7g) in the non-glacier basin. The corrections show larger variation in the westerlies-dominated basins than in the monsoon-dominated basins. The corrections greatly increase the precipitation estimates by about 150%–300% in the glacier area of the westerlies-dominated upper Indus, Amu and Syr Darya basins, and largely decrease the precipitation estimates by about 75%–100% in downstream Yarkant, Amu Darya and Syr Darya basins. In the monsoon-dominated basins, the corrected precipitation increases the estimates by about 50%–130% in the glacier area in the Yarlung Zangbo basin, and decrease the precipitation estimates by about 15%–60% in the other basins.

4.2 Hydrological evaluations of the corrected precipitation

Hydrological models provide a useful tool to inversely evaluate gridded precipitation in streamflow simulations against streamflow observations (Su et al., 2008; Tong et al., 2014a). In this section, the VIC-Glacier model is driven by the daily ERA5_cor precipitation for 1980–2010 in seven selected basins (the UYA, UYE, ULC, UNJ, YZ, UI and UYK) where observed streamflow is available (Fig. 8). For comparison purposes, the simulations with the original ERA5 precipitation data are also included in Fig. 8.

More than 60% of the observed annual total flow occurs in June–September with a single peak in all the selected basins (Fig. 8), except for the UYE basin (Fig. 8b), where the flow regime is characterized by double peaks with the first one in July and the second one in September. Simulated streamflow with the ERA5_cor precipitation estimates more successfully reproduce seasonal patterns of the observed streamflow (Fig. 8, Table A1), with the NSE of 0.7–0.9 and the RB of within ±6%. The ERA5_cor eliminates severe overestimations by 44%–

105% in the simulations with the original ERA5 precipitation. For the monsoon-dominated basins (Fig. 8a–e), compared with the observed streamflow, the RB is reduced from 48%–108% in the simulations with the ERA5 to -5%–3% in the simulations with the ERA5_cor precipitation. For the westerlies-dominated UI and UYK basins (Fig. 8f–g), although both the simulations with ERA5_cor and ERA5 precipitation estimates match well with the observations in June–August, the ERA5_cor eliminates severe overestimations by 20%–50% in the simulated streamflow forced by the original ERA5 in October–May. The streamflow simulations improvements with the ERA5_cor precipitation estimates, relative to those driven by the original ERA5 precipitation estimates, inversely demonstrate the reasonableness of using the corrected ERA5 precipitation as input for hydrological models in the TP basins.

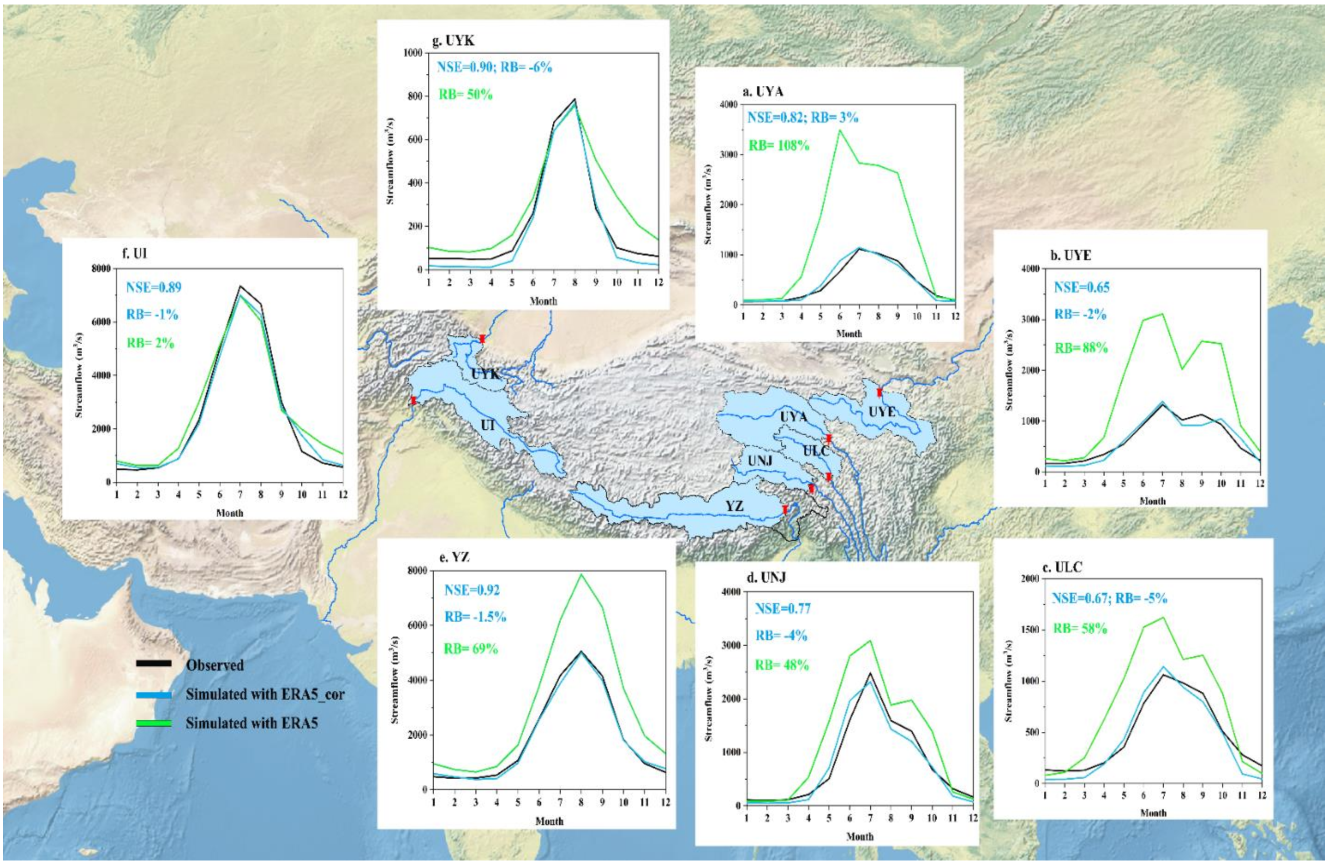

Fig. 8. Observed and the VIC-Glacier model simulated mean monthly total streamflow driven by the ERA5_cor and ERA5 precipitation in the upper regions of Yangtze (UYA), Yellow (UYE), Lancang (ULC), Nujiang (UNJ), Yarlung Zangbo (YZ), Indus (UI), and Yarkant (UYK) river basins for 1980–2010, respectively.

Based on the good results driven by ERA5_cor precipitation (Fig. 8), contributions of rainfall, snowmelt, and glacier runoff to the total river flows are further quantified in the selected basins in the TP (Fig. 9). Dominant water sources are inconsistent among these basins. In the monsoon-dominated UYA, UYE, ULC, UNJ, and YZ basins (Fig. 9a–e), rainfall runoff contribution ranges from 64%–78% among the basins, suggesting that monsoon rainfall plays an important role over these basins. Glacier runoff contributes about 1%–14% to total runoff, with the most in the YZ basin (14%) and the least in the UYE basin (1%). This result is consistent with the glacier area for monsoon-dominated basins (Table 1). Snowmelt runoff contributes about 16–25% to the total runoff in these five monsoon-dominated basins (Fig. 9a–e).

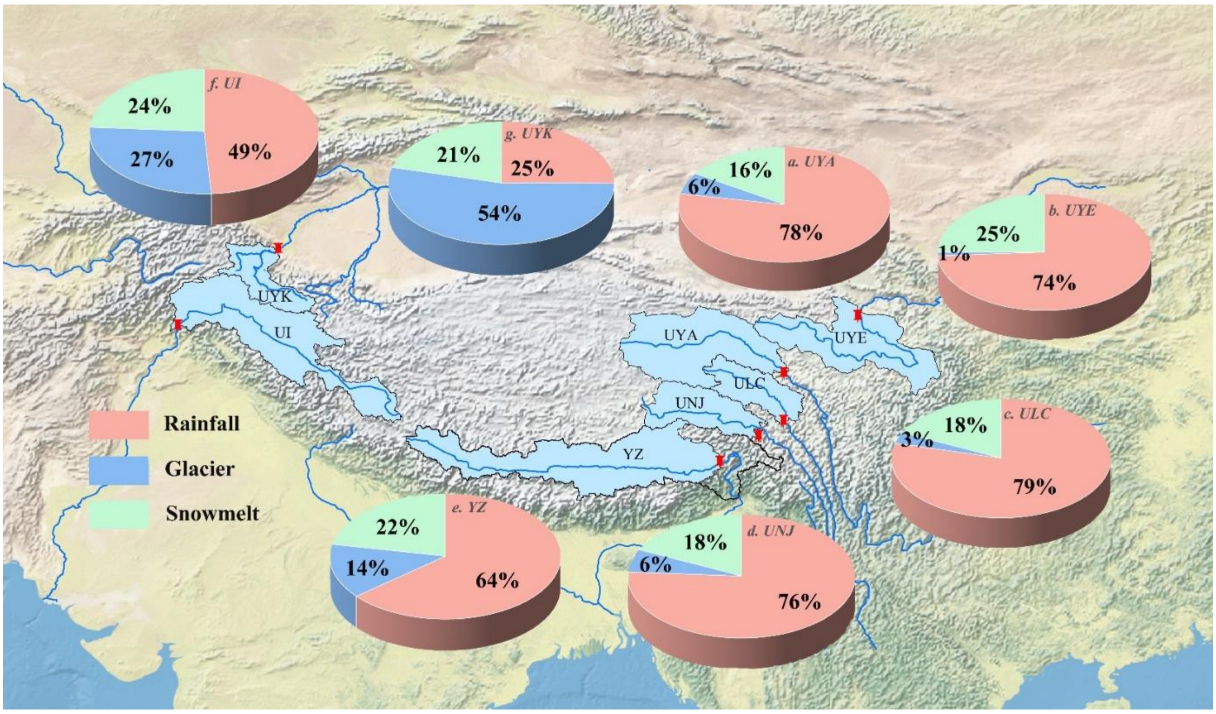

Fig. 9. Contributions of rainfall, snowmelt, and glacier runoff to total annual runoff for seven basins in the Third Pole for 1980–2010, respectively.

In the westerlies-dominated UI and UYK basins with larger glacier cover, the importance of glacier runoff is greater than in the monsoon-dominated basins. However, differences of dominant water source are presented between these two basins (Fig. 9f, g). In the UI basin, the contributions of rainfall, glacier and snowmelt runoff are about 49%, 27%, and 24%, respectively (Fig. 9f). Contributions of rainfall, glacier and snowmelt runoff are about 25%, 54%, and 21% in the UYK basin, respectively (Fig. 9g), suggesting glacier runoff is a dominant water source.

5 Discussion

5.1 Precipitation uncertainties for melt runoff contribution

The accuracy of relative contribution of meltwater to runoff is mostly influenced by precipitation estimates. The overestimation/underestimation of precipitation would be compensated by an underestimate/overestimate of glacier runoff in the model simulation. Sun and Su (2020) suggested that the contribution of glacier runoff would increase about 7–10% with decrease about 20% of mean annual precipitation.

Table 4 summarizes relevant studies on simulated runoff contributions in the seven TP basins, with either different hydrological models or different precipitation datasets. For the monsoon-dominated basins where glaciers only cover about 0.1%–1.5% (Table 1), total runoff is mostly dominated by rainfall. Zhang et al. (2013) and Zhao et al. (2019) simulated runoff components with interpolated CMA gridded precipitation estimates by the VIC-Glacier model in the monsoon-dominated basins, and suggested that glacier runoff contributed about 0.4%–14% to the total runoffs, with the largest in the YZ (6%–14%) and the least in the UYE basin (0.4%–0.8%). However, Khanal et al. (2021) estimated that glacier runoff only contributed about 0.1%–1.8% to the total runoffs, which is mostly due to the large overestimation of the ERA5 precipitation (RB of 50%–280%). These differences in meltwater contribution calculations mostly resulted from the large uncertainties in precipitation estimates used for hydrological models. Aiming to address the “real” precipitation magnitude in the YZ basin, Sun and Su (2020) estimated a basin-wide mean annual precipitation of 544 mm for 1961–2016 in the upper stream of the Nuxia hydrological station in the YZ basin based on the reconstructed precipitation data. It estimated that glacier runoff contributed about 14% to total runoff (Table 4). In this study, we also estimated that glacier runoff contributed about 14% to total runoff in the Nuxia hydrological station of the YZ basin, which is mostly due to the similar mean annual precipitation estimates (560 mm).

These differences in meltwater contributions resulting from precipitation estimates are more obvious in the westerlies-dominated upper Indus, due to its large glacier percentage (about

12%, Table 1). Lutz et al. (2014) simulated streamflow with APHRODITE precipitation estimates (346 mm) by the Spatial Processes in Hydrology (SPHY) model, and suggested glacier runoff contribution of about 40.6% in the UI basin. With the improvement of precipitation estimates (681 mm) in Lutz et al. (2016), glacier runoff contribution is reduced. Using the same hydrological model, Khanal et al. (2021) estimated glacier runoff contribution of about 5.1% with the ERA5 precipitation estimates (832 mm), which is less than the results in Lutz et al. (2014) and Lutz et al. (2016). Mukhopadhyay and Khan (2014) and Mukhopadhyay and Khan (2015) used statistical and hydrograph separation approaches to calculate the contribution of glacier runoff in the UI, and suggested that glacier runoff contributed about 21%–26% to total runoff. In our new work (*Su et al. 2021, manuscript submitted to Earth's Future*), the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) precipitation estimates was corrected for flow simulations in the UI basin, and it estimated that glacier runoff contributed about 24% to total runoff, with mean annual precipitation estimates of 629 mm. In this study, we estimated glacier runoff contribution of about 27%, which is similar to Mukhopadhyay and Khan (2014; 2015) and our new work.

In summary, the uncertainties of precipitation estimates result in large differences in streamflow simulations and quantifying meltwater contributions in glacier-affected basins. Based on the good simulated results with corrected ERA5 precipitation estimates, and in comparison with other studies, our simulated results may represent the most reliable estimation and the best understanding of runoff contributions thus far in the selected TP basins.

Table 4. Summary of relevant studies on simulated runoff component contributions in selected basins of the Third Pole.

Basin	Hydrological station	Runoff contribution (%)			Period	Method	Precipitation		References
		Glacier	Snowmelt	Rainfall			Data	Mean annual estimates (mm)	
UYA	Zhimenda	6.5	22.2	71.3	1963–2005	VIC+DD	Interpolated CMA data	333	Zhang et al. (2013)
		3.7	12.2	84.1	1971–2010	VIC+DD	Interpolated CMA data	—	Zhao et al. (2019)
		0.2	5.5	71	1985–2014	SPHY+DD	ERA5	1127	Khanal et al. (2021)
		6	***16***	***78***	***1980–2000***	***VIC+DD***	***Corrected ERA5***	***383***	***This Study***
UYE	Tangnaihai	0.8	22.4	76.8	1961–1999	VIC+DD	Interpolated CMA data	515	Zhang et al. (2013)
		0.35	15.3	84.35	1971–2010	VIC+DD	Interpolated CMA data	—	Zhao et al. (2019)
		0.1	9.6	63.9	1985–2014	SPHY+DD	ERA5	751	Khanal et al. (2021)
		1	***25***	***74***	***1980–2000***	***VIC+DD***	***Corrected ERA5***	***528***	***This Study***
ULC	Changdu	1.4	20.9	77.7	1961–2000	VIC+DD	Interpolated CMA data	527	Zhang et al. (2013)
		0.9	32.5	43.9	1998–2007	SPHY+DD	APHRODITE	642	Lutz et al. (2014)
		1.3	28.8	69.9	1971–2010	VIC+DD	Interpolated CMA data	—	Zhao et al. (2019)
		0.3	7.4	55.1	1985–2014	SPHY+DD	ERA5	1066	Khanal et al. (2021)
		3	***18***	***79***	***1980–2000***	***VIC+DD***	***Corrected ERA5***	***570***	***This Study***
UNJ	Jiayuqiao	4.8	20.4	74.8	1980–1985	VIC+DD	Interpolated CMA data	607	Zhang et al. (2013)
		8.3	27.5	42	1998–2007	SPHY+DD	APHRODITE	595	Lutz et al. (2014)
		4.4	28.3	67.3	1971–2010	VIC+DD	Interpolated CMA data	—	Zhao et al. (2019)
		7.3	6.9	85.8	1964–2013	VIC+DD	Interpolated CMA data	—	Zhang et al. (2020)

		1.4	14.7	55	1985–2014	SPHY+DD	ERA5	1091	Khanal et al. (2021)
		6	***18***	***76***	***1980–2000***	***VIC+DD***	***Corrected ERA5***	***614***	***This Study***
		11.6	23	65.4	1961–1999	VIC+DD	Corrected CMA data	540	Zhang et al. (2013)
		16	9	59	1998–2007	SPHY+DD	APHRODITE	573	Lutz et al. 2014
		15	27.3	57.7	1971–2000	VIC+DD	Corrected CMA data	540	Su et al. (2016)
		9.9	10.6	79.5	2003–2014	CREST	CGDPA, TMPA	—	Chen et al. (2017)
YZ	Nuxia	5.5	23.1	71.4	1971–2010	VIC+DD	Interpolated CMA data	—	Zhao et al. (2019)
		13.9	23.8	62.3	1980–2000	VIC+DD	Reconstructed gridded data	544	Sun and Su (2020)
		1.8	13.2	62.1	1985–2014	SPHY+DD	ERA5	2018	Khanal et al. (2021)
		18.4	22	69.6	2001-2010	isoGSM	CMFD	—	Nan et al. (2021)
		14	***22***	***64***	***1980–2000***	***VIC+DD***	***Corrected ERA5***	***560***	***This Study***
		48.2	31	20.8	1969–1997	VIC+DD	APHRODITE	425	Zhang et al. (2013)
		40.6	21.8	37.6	1998–2007	SPHY+DD	APHRODITE	346	Lutz et al. (2014)
		21	49	30	1962–2010	Statistical analyses by GIS	—	—	Mukhopadhyay and Khan (2014
UI	Besham	26	44	30	1969–2010	Hydrograph separation	—	—	Mukhopadhyay and Khan (2015
		55		45	1976–2007	SPHY+DD	Corrected APHRODITE	681	Lutz et al. (2016)
		5.1	39.7	43.9	1985–2014	SPHY+DD	ERA5	832	Khanal et al. (2021)
		27	***24***	***49***	***1980–2000***	***VIC+DD***	***Corrected ERA5***	***628***	***This Study***
		52	26	22	1960-2015	VIC+DD	Corrected gauge-based data	234	Kan et al. (2018)
UYK	Kaqun	52	17	16	1968-2007	Glacier-enhanced SWAT	Corrected APHRODITE	—	Wang et al. (2016)

	41	24	35	1966-1995	Glacier-enhanced SWAT	Corrected APHRODITE	—	Luo et al. (2018)
	48	~20	~14	1965-2007	Glacier-enhanced SWAT	Corrected APHRODITE	212	Wang et al. (2021)
	54	***21***	***25***	***1980–2000***	***VIC+DD***	***Corrected ERA5***	***244***	***This Study***

Note: UYA=upper Yangtze; UYE=upper Yellow; ULC=upper Lancan; UNJ=upper Nujiang; UI=upper Indus; YZ=Yarlung Zangbo; VIC+DD=The Variable Infiltration Capacity (VIC) linked with a degree-day glacier melting model; SPHY+DD=The Spatial Processes in Hydrology (SPHY) linked with a degree-day glacier melting model; CREST=Coupled Routing and Excess Storage model; SWAT = Soil and Water Assessment Tool; isoGSM=Scripps global spectral model with water isotopes incorporated; CMFD= China Meteorological Forcing Dataset.

5.2 Uncertainties and limitations of the RF algorithm

The results of this study heavily rely on the RF algorithm and are therefore subject to some limitations associated with the complex terrain and climate controls as well as scarce observations. Reasonable forcing inputs and parameters are crucial for the training and test of the RF algorithm.

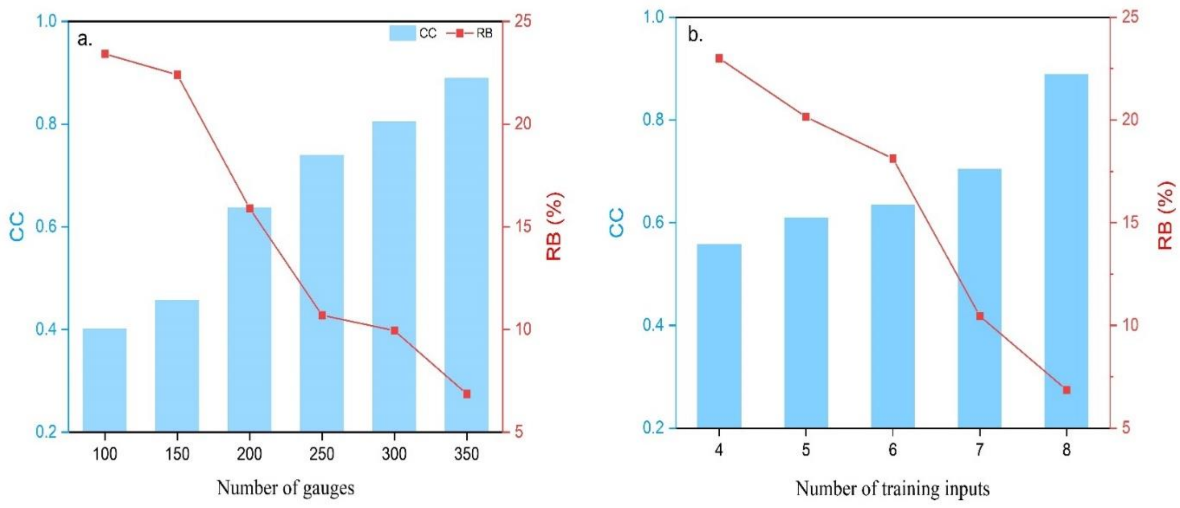

Fig. 10. Sensitivity of the RF algorithm to the numbers of precipitation gauge and selected training inputs in terms of the CC and RB (%).

The limited availability of gauge precipitation observation is an important factor for uncertainties in the results. Although we try our best to improve observation basis by installing new gauges or collecting existing gauge observations, the inadequate coverage and difficulty of maintaining monitoring networks is still the main issue for precipitation evaluation and correction. Fig. 8 shows the sensitivity of the RF algorithm to the numbers of gauge observations and forcing inputs. The observations from 360 gauges in the westerlies-dominated basins are selected as an example. With the gauge numbers increasing from 100 to 350 for the RF training, the CC between simulations by the RF and observations from the rest of gauge in the testing set increases from 0.4 to 0.9, and the RB decreases from 23% to 6% (Fig. 10a). This is similar to forcing data that we selected for the RF algorithm. The numbers of forcing inputs are randomly selected from 4 to 8, the CC increases from 0.56 to 0.87, and the RB decreases from 23% to 6% (Fig. 10b). It is therefore essential for the training of the RF algorithm to explore more related and reasonable input features. Improvements of the performance of the RF algorithm could benefit from involving more gauge observations in the future, even though our results demonstrate that the current used gauges serve as a good data base for the implementation of the RF-correction procedure for the ERA5 estimates.

Parameters themselves have their own uncertainties, which are ideally all taken into account. The *mtry*, *node_size* and *sample_fraction* have been identified as the most sensitive parameters in the RF algorithm (Adhikari et al., 2020); sensitivity of the RF performance to such parameters in terms of the OOB MSE is then shown in Fig. 11. The OOB MSE decreases when *mtry* increases from 1 to 3, while increasing when *mtry* increases from 3 to 8 (Fig. 11a). The OOB MSE increases with increasing *node_size* (Fig. 11b), while decreasing with increasing *sample_fraction* (Fig. 11c). To reduce the uncertainty originating from the parameters, the RF algorithm has been trained in multiple runs and a grid search has been run to optimize the model parameters.

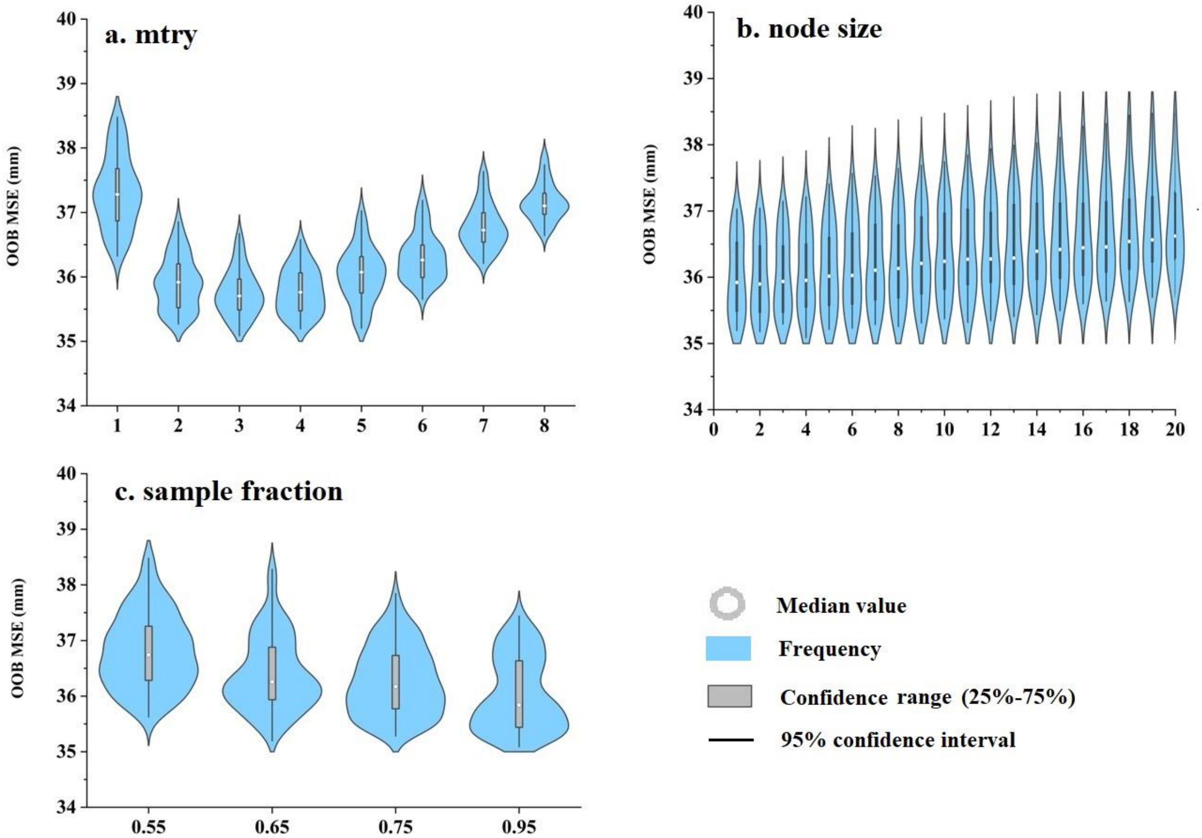

Fig. 11. Sensitivity of the RF algorithm to parameters of the *mtry*, *node size* and *sample fraction* in terms of out-of-bag mean squared error (OOB MSE).

6 Conclusions

In this study, the original ERA5 precipitation estimate is corrected based on 940 rain gauges, WRF-9km gridded estimates and the RF algorithm for 11 upper basins in the TP, including the monsoon-dominated UYA, UYE, ULC, UNJ, and YZ basins, and westerlies-dominated UI, UAMD, USRD, UYK, UHT and UAKS basins. The corrected dataset is evaluated by gauge observations at point scale, and inversely evaluated by the VIC-Glacier hydrological model. The main findings are summarized below:

1. A gridded daily precipitation dataset with a spatial resolution of 10 × 10 km for 1951–2020 is constructed by corrections of ERA5 precipitation for 11 basins in the TP.

2. The corrected ERA5 (ERA5_cor) precipitation estimates agree well with gauge observations with CC of 0.7–0.8 and RB of 5–9% by eliminating severe overestimations in the original ERA5. The corrections greatly decrease precipitation estimates by 10%–50% among the 11 basins from a mean annual precipitation of 402–1267 mm in the original ERA5 to 244–768 mm in the ERA5_cor for 1951–2020. The corrections show more detailed information on the spatial pattern than that in the original ERA5 estimates.

3. Simulated streamflow with the ERA5_cor estimates successfully reproduces seasonal patterns of observed streamflow in seven upper basins of the TP, with the NSE of 0.7–0.9 and RB of within ±6%, demonstrating the accuracy of the corrected ERA5 precipitation estimates as input for hydrological models. Based on these well simulated results, we provide updated estimates of runoff component contributions in the basins, with a dominant water source of rainfall runoff (64%–78%) in five monsoon-dominated basins and glacier runoff playing a more important role in westerlies-dominated basins than the monsoon-dominated basins.

CRediT authorship contribution statement

He Sun: Conceptualization, Formal analysis, Investigation, Methodology, Resources, Visualization, Writing draft. Tandong Yao and Fengge Su: Conceptualization, Resources, Visualization, Funding acquisition, Writing- review & editing. Zhihua He and Guoqiang Tang: Formal analysis and Writing (review and editing); Ning Li, Bowen Zheng, Jingheng Huang and Fanchong Meng: Formal analysis, Methodology, Writing-review & editing. Tinghai Ou and Deliang Chen: Writing (review and editing) and providing WRF-9km data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was financially supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDA20100300), and the Second Tibetan Plateau Scientific Expedition and Research Program (2019QZKK0201). We acknowledge that the Figure 1 in this study was previously published under the terms of the Creative Commons Attribution 4.0 license in Sun, H., Su, F., Yao, T. et al., 2021. General overestimation of ERA5 precipitation in flow simulations for High Mountain Asia basins. Environ. Res. Commun. 3(12): 121003. https://doi.org/10.1088/2515-7620/ac40f0.

Data Availability Statement

The constructed gridded daily precipitation dataset (10 km × 10 km) for 1951–2020 in 11 TP basins in this study can be downloaded from the National Tibetan Plateau Data Center (https://data.tpdc.ac.cn). The daily WRF-9km precipitation data is available from the Regional Climate Group, Department of Earth Sciences, University of Gothenburg (http://biggeo.gvc.gu.se/TPReanalysis/). ERA5 data is downloaded from https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5. Station data of precipitation are from the China Meteorological Administration (CMA; http://cdc.cma.gov.cn), the Global Historical Climatology Network (GHCN, https://www.ncdc.noaa.gov/ghcn-monthly) and the Pakistan Meteorological Department (https://www.pmd.gov.pk/en/).

References

Adhikari, A., Ehsani, M.R., Song, Y. et al., 2020. Comparative assessment of snowfall retrieval from microwave humidity sounders using machine learning methods. Earth Space Sci. 7(11). https://doi.org/10.1029/2020ea001357

Bookhagen, B., Burbank, D.W., 2010. Toward a complete Himalayan hydrological budget: Spatiotemporal distribution of snowmelt and rainfall and their impact on river discharge. J. Geophys. Res. Atmos. 115(F3). https://doi.org/10.1029/2009jf001426

Chen, X., Long, D., Hong, Y. et al., 2017. Improved modeling of snow and glacier melting by a progressive two-stage calibration strategy with GRACE and multisource data: How snow and glacier meltwater contributes to the runoff of the Upper Brahmaputra River basin? Water Resour. Res. 53: 2431–2466. https://doi.org/10.1002/2016WR019656

Cuo, L., Zhang, Y., 2017. Spatial patterns of wet season precipitation vertical gradients on the Tibetan Plateau and the surroundings. Sci. Rep. 7(1): 5057. https://doi.org/10.1038/s41598-017-05345-6

Dahri, Z.H., Ludwig, F., Moors, E. et al., 2021. Spatio- Temporal Evaluation of Gridded Precipitation Products for the High- Altitude Indus Basin. Int. J. Climatol. 41:4283–4306. https://doi.org/10.1002/joc.7073

Ehsani, M.R., Arevalo, J., Risanto, C.B. et al., 2020. 2019–2020 Australia Fire and Its Relationship to Hydroclimatological and Vegetation Variabilities. Water 12(11). https://doi.org/10.3390/w12113067

Gao, Y., Chen, F., Jiang, Y., 2020. Evaluation of a convection-permitting modeling of precipitation over the Tibetan Plateau and its influences on the simulation of snow-cover fraction. J. Hydrometeorol. 21(7): 1531–1548. https://doi.org/10.1175/jhm-d-19-0277.1

Gao, Y., Liu, M., 2013. Evaluation of high-resolution satellite precipitation products using rain gauge observations over the Tibetan Plateau. Hydrol. Earth Syst. Sci. 17(2): 837–849. https://doi.org/10.5194/hess-17-837-2013

He, Z., Duethmann, D., Tian, F., 2021. A meta-analysis based review of quantifying the contributions of runoff components to streamflow in glacierized basins. 603: 126890. J. Hydrol. https://doi.org/10.1016/j.jhydrol.2021.126890

Hersbach, H., Bell, B., Berrisford, P. et al., 2020. The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society. 146(730): 1999–2049. https://doi.org/10.1002/qj.3803

Hock, R., 2003. Temperature index melt modelling in mountain areas. J. Hydrol. 282(1–4): 104-115. https://doi.org/10.1016/s0022-1694(03)00257-9

Huang, J., Su, F., 2019. Hydrological response of the upper Syr Darya basin to climate change, American Geophysical Union Fall Meeting, San Francisco, pp. GC42B–06.

Janitza, S., R. Hornung, 2018. On the overestimation of random forest's out-of-bag error. PloS One 13(8): e0201904. https://doi.org/10.1371/journal.pone.0201904

Jiang, Y., Yang, K., Shao, C. et al., 2021. A downscaling approach for constructing a high-resolution precipitation dataset over the Tibetan Plateau from ERA5 reanalysis. Atmos Res. 256: 105574. https://doi.org/10.1016/j.atmosres.2021.105574

Kan, B., Su, F., Xu, B. et al., 2018. Generation of High Mountain Precipitation and Temperature Data for a Quantitative Assessment of Flow Regime in the Upper Yarkant Basin in the Karakoram. J. Geophys. Res. Atmos. 123(16): 8462–8486. https://doi.org/10.1029/2017jd028055

Khanal, S., Lutz, A., Kraaijenbrink, P. et al., 2021. Variable 21st century climate change response for rivers in High Mountain Asia at seasonal to decadal time scales. Water Resour. Res. 57(5): e2020WR029266. https://doi.org/https://doi.org/10.1029/2020WR029266

Lai, H.-W., H. W. Chen, J. Kukulies, T. Ou, D. Chen, 2021. Regionalization of seasonal precipitation over the Tibetan Plateau and associated large-scale atmospheric systems. J Climate. 34(7): 2635–2651. DOI: DOI: 10.1175/JCLI-D-20-0521.1

Li, C., 2019. Study on discharge simulation and projection over the watersheds of Upper Indus

Basin. Doctor’s thesis. University of Chinese Academy of Sciences.

Liang, X., Lettenmaie, D.P., Wood, E.F. et al., 1994. A simple hydrologically based model of land-surface water and energy fluxes. J. Geophys. Res. Atmos. 99:(D7): 14415–14428. https://doi.org/Doi 10.1029/94jd00483

Liang, X., Lettenmaier, D.P., Wood, E.F., 1996. One-dimensional statistical dynamic representation of subgrid spatial variability of precipitation in the two-layer variable infiltration capacity model. J. Geophys. Res. Atmos. 101(D16): 21403–21422. https://doi.org/10.1029/96jd01448

Luo, Y., Wang, X., Piao, S. et al., 2018. Contrasting streamflow regimes induced by melting glaciers across the Tien Shan-Pamir-North Karakoram. Sci. Rep. 8(1): 16470. https://doi.org/10.1038/s41598-018-34829-2

Lutz, A.F., Immerzeel, W.W., Kraaijenbrink, P.D. et al., 2016. Climate Change Impacts on the Upper Indus Hydrology: Sources, Shifts and Extremes. PLoS One 11(11): e0165630. https://doi.org/10.1371/journal.pone.0165630

Lutz, A.F., Immerzeel, W.W., Shrestha, A.B. et al., 2014. Consistent increase in High Asia's runoff due to increasing glacier melt and precipitation. Nat. Clim. Chang. 4(7): 587–592. https://doi.org/10.1038/nclimate2237

Ma, Y., Zhang, Y., Yang, D. et al., 2015. Precipitation bias variabilityversusvarious gauges under different climatic conditions over the Third Pole Environment (TPE) region. Int. J. Climatol. 35(7): 1201–1211. https://doi.org/10.1002/joc.4045

Mölg, T., Maussion, F., Scherer, D., 2013. Mid-latitude westerlies as a driver of glacier variability in monsoonal High Asia. Nat. Clim. Chang. 4(1): 68–73. https://doi.org/10.1038/nclimate2055

Mukhopadhyay, B., Khan, A., 2014. A quantitative assessment of the genetic sources of the hydrologic flow regimes in Upper Indus Basin and its significance in a changing climate. J. Hydrol. 509: 549–572. https://doi.org/10.1016/j.jhydrol.2013.11.059

Mukhopadhyay, B., Khan, A., 2015. A reevaluation of the snowmelt and glacial melt in river flows within Upper Indus Basin and its significance in a changing climate. J. Hydrol. 527:

119–132. https://doi.org/10.1016/j.jhydrol.2015.04.045

Nan, Y., He, Z., Tian, F. et al., 2021. Can we use precipitation isotope outputs of isotopic general circulation models to improve hydrological modeling in large mountainous catchments on the Tibetan Plateau? Hydrol. Earth Syst. Sci. 25(12): 6151-6172. https://doi.org/10.5194/hess-25-6151-2021

Oppel, H., Fischer, S., 2020. A New Unsupervised Learning Method to Assess Clusters of Temporal Distribution of Rainfall and Their Coherence with Flood Types. Water Resour. Res. 56(5). https://doi.org/10.1029/2019wr026511

Ou, T., Chen, D., Chen, X. et al., 2020. Simulation of summer precipitation diurnal cycles over the Tibetan Plateau at the gray-zone grid spacing for cumulus parameterization. Clim. Dynam. 54(7): 3525–3539. https://doi.org/10.1007/s00382-020-05181-x

Su, F., Hong, Y., Lettenmaier, D.P., 2008. Evaluation of TRMM Multisatellite Precipitation Analysis (TMPA) and Its Utility in Hydrologic Prediction in the La Plata Basin. J. Hydrometeorol. 9(4): 622–640. https://doi.org/10.1175/2007jhm944.1

Su, F., Zhang, L., Ou, T. et al., 2016. Hydrological response to future climate changes for the major upstream river basins in the Tibetan Plateau. Glob. Planet. Change. 136: 82–95. https://doi.org/10.1016/j.gloplacha.2015.10.012

Sun, H., Su, F., 2020. Precipitation correction and reconstruction for streamflow simulation based on 262 rain gauges in the upper Brahmaputra of southern Tibetan Plateau. J. Hydrol. 590: 125484. https://doi.org/10.1016/j.jhydrol.2020.125484

Sun, H., Su, F., He, Z. et al., 2021a. Hydrological evaluation of high-resolution precipitation estimates from the WRF model in the Third Pole river basins. J. Hydrometeorol. 22(8): 2055–2071. https://doi.org/10.1175/jhm-d-20-0272.1

Sun, H., Su, F., Huang, J. et al., 2020. Contrasting precipitation gradient characteristics between westerlies and monsoon dominated upstream river basins in the Third Pole. Chin. Sci. Bull. 65(1): 91–104. https://doi.org/10.1360/tb-2019-0491

Sun, H., Su, F., Yao, T. et al., 2021b. General overestimation of ERA5 precipitation in flow simulations for High Mountain Asia basins. Environ. Res. Commun. 3(12): 121003.

https://doi.org/10.1088/2515-7620/ac40f0

Tang, G., Clark, M.P., Papalexiou, S.M. et al., 2020. Have satellite precipitation products improved over last two decades? A comprehensive comparison of GPM IMERG with nine satellite and reanalysis datasets. Remote Sens. Environ. 240: 111697. https://doi.org/10.1016/j.rse.2020.111697

Tong, K., Su, F., Xu, B., 2016. Quantifying the contribution of glacier meltwater in the expansion of the largest lake in Tibet. J. Geophys. Res. Atmos. 121(19): 11158–11173. https://doi.org/10.1002/2016jd025424

Tong, K., Su, F., Yang, D. et al., 2014a. Evaluation of satellite precipitation retrievals and their potential utilities in hydrologic modeling over the Tibetan Plateau. J. Hydrol. 519: 423–437. https://doi.org/10.1016/j.jhydrol.2014.07.044

Tong, K., Su, F., Yang, D. et al., 2014b. Tibetan Plateau precipitation as depicted by gauge observations, reanalyses and satellite retrievals. Int. J. Climatol. 34(2): 265–285. https://doi.org/10.1002/joc.3682

Viviroli, D., Archer, D.R., Buytaert, W. et al., 2011. Climate change and mountain water resources: overview and recommendations for research, management and policy. Hydrol. Earth Syst. Sci. 15(2): 471–504. https://doi.org/10.5194/hess-15-471-2011

Wang, Q., Huang, J., Liu, R. et al., 2020. Sequence-based statistical downscaling and its application to hydrologic simulations based on machine learning and big data. J. Hydrol. 586. https://doi.org/10.1016/j.jhydrol.2020.124875

Wang, X., Luo, Y., Sun, L. et al., 2016. Attribution of Runoff Decline in the Amu Darya River in Central Asia during 1951–2007. J. Hydrometeorol. 17(5): 1543–1560. https://doi.org/10.1175/jhm-d-15-0114.1

Wang, X., Luo, Y., Sun, L. et al., 2021. Different climate factors contributing for runoff increases in the high glacierized tributaries of Tarim River Basin, China. J. Hydrol.-Reg. Stud. 36: 100845. https://doi.org/10.1016/j.ejrh.2021.100845

Yao T. 2014. TPE international program: a program for coping with major future environmental challenges of The Third Pole region. Progress in Geography. 33(7): 884-892.

https://doi.org/10.11820/dlkxjz.2014.07.003
Yang, D., Kane, D., Zhang, Z. et al., 2005. Bias corrections of long-term (1973–2004) daily precipitation data over the northern regions. Geophys. Res. Lett. 32(19). https://doi.org/10.1029/2005gl024057
You, Q., Fraedrich, K., Ren, G. et al., 2012. Inconsistencies of precipitation in the eastern and central Tibetan Plateau between surface adjusted data and reanalysis. Theor. Appl. Climatol. 109(3): 485–496. https://doi.org/10.1007/s00704-012-0594-1
Zhang, L., Su, F., Yang, D. et al., 2013. Discharge regime and simulation for the upstream of major rivers over Tibetan Plateau. J. Geophys. Res. Atmos. 118(15): 8500–8518. https://doi.org/10.1002/jgrd.50665
Zhang, Y., Ye, A., 2021. Machine learning for precipitation forecasts post-processing —Multi-model comparison and experimental investigation. J. Hydrometeorol. 22(11): 3065-3085. https://doi.org/10.1175/jhm-d-21-0096.1
Zhang, Y., Xu, C.-Y., Hao, Z. et al., 2020. Variation of Melt Water and Rainfall Runoff and Their Impacts on Streamflow Changes during Recent Decades in Two Tibetan Plateau Basins. Water 12(11). https://doi.org/10.3390/w12113112
Zhao, Q., Ding, Y., Wang, J. et al., 2019. Projecting climate change impacts on hydrological processes on the Tibetan Plateau with model calibration against the Glacier Inventory Data and observed streamflow. J. Hydrol. 573: 60–81. https://doi.org/10.1016/j.jhydrol.2019.03.043