How do GPM IMERG precipitation estimates perform as hydrological model forcing? Evaluation for 300 catchments across Mainland China

Jiang, Liguang; Bauer-Gottwein, Peter

Published in:
Journal of Hydrology

Link to article, DOI:
[10.1016/j.jhydrol.2019.03.042](https://doi.org/10.1016/j.jhydrol.2019.03.042)

Publication date:
2019

Document Version
Peer reviewed version

[Link back to DTU Orbit](#)

Citation (APA):
Jiang, L., & Bauer-Gottwein, P. (2019). How do GPM IMERG precipitation estimates perform as hydrological model forcing? Evaluation for 300 catchments across Mainland China. *Journal of Hydrology*, 572, 486-500. <https://doi.org/10.1016/j.jhydrol.2019.03.042>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

**How do GPM IMERG precipitation estimates perform as
hydrological model forcing? Evaluation for 300 catchments
across Mainland China**

Liguang Jiang*, Peter Bauer-Gottwein

*Department of Environmental Engineering, Technical University of Denmark,
2800 Kgs. Lyngby, Denmark*

*Corresponding author at: Room 247, building 115, Bygningstorvet, DTU
Environment, 2800 Kgs. Lyngby, Denmark

E-mail address: ljia@env.dtu.dk (L. Jiang)

Abstract

Accurate precipitation estimates are essential for reliable hydrological simulations of fluxes and states. Satellite precipitation products have been widely used for hydrological applications, especially for ungauged regions. The Global Precipitation Measurement (GPM) product is starting to provide a new generation of precipitation estimates. This work first evaluates several GPM Integrated Multi-satellite Retrievals (IMERG) against gauge-based precipitation over Mainland China for a two-year period (2016 - 2017), and then explores the feasibility of using GPM to force a lumped hydrological model over 300 catchments of varying size and climate.

The analysis indicates that IMERG Final run estimates (gauge-adjusted) agree well with gauge-based precipitation at daily scale, while IMERG Early run estimates (near-real time without gauge adjustments) are consistently lower than the rain gauge records. Nevertheless, both Final run and Early run estimates are slightly better than TRMM 3B42 estimates in terms of rainfall detection skill. Moreover, all three products show better skill over humid regions than over high elevation zones.

When used as hydrological model forcing, IMERG Final and Early estimates provide comparable performances to gauge-based precipitation. However, TRMM 3B42 performs relatively poor in terms of hydrological simulation. Moreover, models generally perform better in humid than in arid areas for all three products. The evaluation sheds light on how IMERG-driven hydrological predictions perform in different regions across China. The findings also indicate the potential of IMERG Early run for flood forecasting and Final run for climatological/hydrological modeling in ungauged or poorly gauged basins.

Keywords

Satellite precipitation products, GPM IMERG, rainfall-runoff modeling, HBV, China

1. Introduction

As a major component of the water cycle and a key input of hydrologic models, precipitation is one of the most important variables in the field of hydrometeorology. Due to the heterogeneity of precipitation and limited spatial density of gauging networks, the integration of satellite precipitation products (SPPs) with hydrologic models is a promising approach to improve the understanding of hydrological processes across continental scales (Mei et al., 2016). There is a plethora of studies, which use satellite precipitation as forcing data for different hydrological applications, especially for ungauged basins. Therefore, the accuracy of precipitation is of great importance for hydrological modeling and forecasting of extreme events in these regions. With recent advances in remote sensing technology, SPPs are becoming more accurate and increase in spatial and temporal resolution (Ciabatta et al., 2016; Maggioni et al., 2016). If such products are sufficiently accurate, they can help improve our capability to predict natural hazards, such as floods, droughts, and landslides, especially in ungauged regions (Ashouri et al., 2016; Zambrano-Bigiarini et al., 2017).

During the last few decades, a number of widely used SPPs were developed, such as the CMORPH (Joyce et al., 2004), PERSIANN (Sorooshian et al., 2000), TMPA (Huffman et al., 2007), SM2RAIN-ASCAT (Brocca et al., 2013), etc. Recently, Beck et al. (2017) published a review on the evaluation of 22 precipitation datasets. They found that there are large differences in estimation accuracy, and their performances are site-specific. Maggioni et al. (2016) also found that topography, seasonality, and climatology play an important role in the SPP's performance, especially for the probability of detection and bias.

Since the launch of the Global Precipitation Measurement (GPM) mission, evaluation and benchmarking of SPPs from GPM have been in focus. One of the GPM products is based on the Integrated Multi-satellitE Retrievals for GPM (IMERG), which got much attention since it potentially delivers the 'best' precipitation estimates (Huffman et al., 2017). Three main

improvements are (a) the increased inclination (65°), which allows for the coverage of important additional climate zones; (2) dual frequencies, which add sensitivity to the detection of light precipitation; and (3) high-frequency channels (165.5 and 183.3 GHz) deployed on the imager, which facilitate sensing of both light and solid precipitation (Huffman et al., 2015). GPM provides precipitation measurements at a higher spatial (from 0.1° to 0.25°) and temporal (from half-hourly to 3-hourly) resolution compared with TMPA. However, as an extension of TMPA, IMERG is still in its infancy. The IMERG algorithm has been updated three times, i.e. V3, V4, and V5 (Huffman et al., 2017). Much work has been done to evaluate the consistency and discrepancies between these products. Liu (2016) assessed the systematic differences between TMPA and IMERG at a global scale. Their results indicated that large systematic differences exist and vary with surface type and precipitation rate. Prakash et al. (2016) compared the capacity of the two products to capture heavy rain across India, and their results revealed that IMERG shows notable improvements over TMPA. Meanwhile, some other research studied the performance of IMERG against gauge measurements, and a few studies investigated its hydrological applications (e.g., Wang et al., 2017; Zubieta et al., 2017). The current studies show that IMERG generally performs well although different error characteristics exist over different regions. Most studies conclude that SPP performance is regionally dependent and can differ widely over diverse landscapes, climate zones, latitude, and elevation, etc. (Maggioni et al., 2016; Wang et al., 2017; Zambrano-Bigiarini et al., 2017). However, previous studies mainly focus on the product quality and not on its suitability for hydrological applications, for example, as forcing input of rainfall-runoff simulation. There is a clear need to understand whether IMERG can be transformed into runoff at catchment scale, and how this simulation performs in regions with diverse climatological regimes.

Some studies evaluated the performance of IMERG using independent gauge observations in China. Guo et al. (2016) studied the performance of IMERG against daily ground measurements over China, and concluded that IMERG performs poorly in winter and over-detects slight rain

events in northwestern China. Xu et al. (2017) validated IMERG and 3B42 over the southern Tibetan Plateau, indicating underestimation of light rain events by IMERG and overestimation by 3B42. Meanwhile, Li et al. (2017) and Tang et al. (2016) assessed the streamflow simulation using IMERG as forcing input and showed promising results in one large catchment in China. Similarly, Wang et al. (2017) evaluated the hydrological utility of the IMERG product over the Beijiang River. He et al. (2017) intercompared the performance of TRMM 3B42 and IMERG and evaluated their hydrological application in the upper Mekong River basin. The results indicate that IMERG is capable of detecting extreme rain events and improves the ability to capture rainfall events with moderate intensities. Besides, IMERG tends to outperform 3B42 in hydrological simulations. However, these studies were conducted for a limited number of catchments, and can therefore not support general conclusions. Moreover, catchments were fairly large (e.g. over 80000 km² of Ganjiang River basin, 78000 km² of upstream of Mekong River basin, 38000 km² of Beijing River basin, etc.), leading to combined rainfall and model uncertainty (Beck et al., 2017). In this context, it is of great value to evaluate the hydrological applications of IMERG for different regions with diverse landscapes, topography, and climatic conditions.

This study aims to evaluate IMERG products and their performance in hydrological simulation across the climatically and topographically diverse mainland of China. The objectives are two-fold: (1) to understand the data accuracy and detection skills of IMERG estimates; (2) to investigate the utility of IMERG estimates for hydrological simulations. To achieve the objectives, we assess daily satellite estimates against gauge-based observations at 838 gauging stations; then we use both satellite estimates and gauge-based observations to force lumped hydrological models of 300 catchments. This work provides insights useful for hydrological application of IMERG products for both algorithm developers and end-users.

2. Study area and Data

2.1. Study area

In this study, 300 small to medium-sized catchments were selected (Fig. 1) based on the following criteria. First, catchments with areas less than 5000 km² were used because observed discharge is expected to be dominated by rainfall-runoff generation instead of routing processes (Beck et al., 2016). Second, mainly headwater catchments were selected, where flow alteration due to abstractions/regulation is minimal. According to the Köppen-Geiger climate classification (Kottek et al., 2006), these catchments are distributed across all five main climate types although around 77% are in temperate climate. The annual precipitation ranges from less than hundred mm in the northwest to over 2000 mm in the southeast. Precipitation is almost invariably concentrated in the summer monsoon season (Xiao et al., 2013). The mean elevations of the selected catchments are also very diverse, ranging from a few meters to over 4000 m amsl (Fig. 1). The watershed boundaries are defined using the HydroBASINS product of the HydroSHEDS database (Lehner and Grill, 2013).

Fig. 1. Catchments distribution across mainland China. Background is the Köppen-Geiger climate classification. The pie chart shows the distribution of catchments across the five main climate zones. The bar chart shows the histogram of catchment elevations.

2.2. Gauge data

Meteorological data (including precipitation, temperature, pressure, relative humidity, wind speed, sunshine hours) at 838 stations were downloaded from the China Meteorological Data Service Center (<http://data.cma.cn/en>). This dataset is at daily scale and used in this study for two purposes: evaluation of satellite precipitation estimates and forcing of hydrological modeling.

We collected daily discharge data from the Hydroinfo website managed by the Ministry of Water Resources of China (<http://xxfb.hydroinfo.gov.cn/ssIndex.html>). The available two-year period (2016-2017) was split into one year for calibration and another for validation of the hydrological simulations.

2.3. Satellite products

GPM IMERG contains precipitation estimates from passive microwave sensors on board various satellites in the GPM constellation, and fills in temporal gaps with infrared fields merged from several geostationary satellites (Huffman et al., 2017; Liu, 2016). Since the first release of IMERG Final run (V03D) in 2015, it has undergone many improvements and the latest release is version V05B, which is available from <https://pmm.nasa.gov/data-access/downloads/gpm> in near-real time, i.e. “early” run (IMERG-E), “late” run (IMERG-L), and post-real-time, i.e. “final” run (IMERG-F). IMERG-E is produced about 4 hr after observation time for operational uses, such as potential flood or landslide forecasting. IMERG-L is produced about 12 hr after observation time for users who work in agricultural drought forecasting and monitoring. In contrast, IMERG-F is created post real time (3 months after observation month) using monthly gauge analysis from the

Global Precipitation Climatology Centre (GPCC). The main difference between IMERG-E and IMERG-F is that IMERG-E has no climatological adjustment while IMERG-F is climatologically adjusted at the monthly scale. Moreover, IMERG-E only has forward propagation (extrapolation) and IMERG-F has both forward and backward propagation (allowing interpolation). A detailed description can be found in Huffman et al. (2015, 2017). All three products are at half-hour temporal resolution and $0.1^\circ \times 0.1^\circ$ spatial resolution (ca. 80~116 km² from the north to the south of our study region). In this study, we consider the V05B IMERG-E and IMERG-F products. Daily precipitation is created by aggregating half-hourly data into daily data. Meanwhile, for comparative purposes, TRMM-3B42 V07 is also used in this study. TRMM-3B42 V07 is also corrected with gauge data and available in 0.25° spatial resolution and 3-hourly temporal resolution covering 50°N to 50°S for the period of 1998-present.

We are aware that both IMERG-F and TRMM-3B42 V07 products are climatologically adjusted using the monthly GPCC gauge analysis product, which is derived from measurements by about 6700 gauge stations over the globe (Huffman et al., 2017, 2007; Tian et al., 2007). However, only 194 stations are used in the GPCC network (Wang et al., 2018), thus, most stations used in this study are independent from GPCC. Moreover, the reason for the use of IMERG-E near-real time product is that it allows to show the improvement of IMERG-F product after bias-correction.

3. Methods

3.1. Meteorological evaluation

Satellite estimates are compared with the gauged precipitation at each gauge station. Specifically, for each rain gauge, the grid cell of satellite products containing that gauge is selected. Satellite estimates are assessed in terms of their ability to represent the daily variability of precipitation by four statistical metrics, i.e., relative error (RE), correlation coefficient (CC), root mean squared error (RMSE), and mean absolute error (MAE). RE assesses the accuracy of satellite

estimates, i.e. over- or underestimation; CC indicates the match of the temporal variability between satellite estimates and gauge observations; RMSE shows the magnitude of the errors, and weights larger errors more than smaller ones; and MAE measures the average magnitude of the errors without considering their signs.

To examine the capability of satellite products to detect rain events, several widely used categorical skill metrics are adopted, including the probability of detection (POD), false alarm ratio (FAR), critical success index (CSI), and Heidke skill score (HSS) (Table 1). These metrics are calculated from a 2×2 contingency matrix composed of four parameters (A, B, C, D), where A is the number of observed events correctly detected; B is the number of events detected but not observed; C is the number of events not detected but observed; and D is the sum of events that are neither detected nor observed. Given that in-situ gauge precipitation data are at daily temporal resolution, these detection metrics are calculated based on daily gauge data and corresponding satellite estimates. The rainfall threshold used is 1 mm/day following Dinku et al. (2008). HSS measures the accuracy of the estimates accounting for matches due to random chance. It is normalized by the total range of possible improvement, therefore HSS can safely be compared among different sites. If HSS is one, the estimate is perfect while zero shows no skill, and a negative number indicates that the estimate is worse than a random prediction. POD, FAR, and CSI are affected by the climatology of the study region while HSS is not affected as much (Dinku et al., 2008).

Table 1 Summary of the statistical metrics used to evaluate the performance of IMERG dataset.

All metrics are calculated based on gauging stations and the corresponding satellite grid cells.

Note that G_i indicates gauge observation and S_i is satellite estimate. Index i represents a time step.

$\bar{G}(\bar{S})$ is the mean value of gauge observations (satellite estimates).

Metric	Abbreviation	Equation	Perfect value
--------	--------------	----------	---------------

Relative error	RE (-)	$\frac{\sum S_i - G_i}{\sum G_i}$	0
Correlation coefficient	CC (-)	$\frac{\sum (S_i - \bar{S})(G_i - \bar{G})}{\sqrt{\sum (S_i - \bar{S})^2} \sqrt{\sum (G_i - \bar{G})^2}}$	1
Root mean square error	RMSE (mm day ⁻¹)	$\sqrt{\frac{\sum (S_i - G_i)^2}{n}}$	0
Mean absolute error	MAE (mm day ⁻¹)	$\frac{\sum S_i - G_i }{n}$	0
Probability of detection	POD (-)	$\frac{A}{A + C}$	1
False alarm ratio	FAR (-)	$\frac{B}{A + B}$	0
Critical success index	CSI (-)	$\frac{A}{A + B + C}$	1
Heidke skill score	HSS (-)	$\frac{2(A * D - B * C)}{((A + C) * (C + D) + (A + B) * (B + D))}$	1

3.2. Hydrological evaluation of areal daily satellite estimates

The hydrological model used in this study is the lumped HBV conceptual model (Bergström, 1995; Kollat et al., 2012). This model is computationally efficient and widely used. The forcing inputs are time series of precipitation, potential evapotranspiration, and temperature. The model consists of four routines, i.e., a degree-day snow routine, a soil moisture routine, a lumped response function, and a routing routine (Fig. 2).

Table 2. Model parameter definition and range

Parameter symbol	Description [Units]	Lower bound	Upper bound
Snow routine			
<i>TS</i>	Threshold temperature [°C]	-2	3
<i>DDF</i>	Degree-Day factor [mm/°C/day]	1	10
<i>CFR</i>	Refreezing factor [-]	0	0.3
<i>WHC</i>	Water holding capacity of snow [-]	0	0.5
Soil routine			
<i>LP</i>	Evapotranspiration limit [-]	0.3	1
<i>FC</i>	Field capacity [mm]	10	550
<i>β</i>	Exponential parameter [-]	1	5
Response routine			
<i>L</i>	Near-surface flow threshold [mm]	0	100

K_q	Near-surface flow recession coefficient [day ⁻¹]	0.05	0.5
K_m	Intermediate flow recession coefficient [day ⁻¹]	0.01	0.3
K_s	Base flow recession coefficient [day ⁻¹]	0.005	0.1
$PERC$	Maximal flow from upper to lower zone [mm/day]	0	5
Routing routine			
$MaxBas$	Flow routing coefficient [day]	0.1	5

191

192 Snow accumulation occurs when temperature is below the threshold temperature while
193 snowmelt starts when temperature is above the threshold. The snowpack retains melt water until a
194 certain level (WHC) is reached, and excess water recharges the soil moisture accounting module.
195 The soil moisture accounting module utilizes a storage capacity distribution function for the storage
196 elements of the catchment. Actual evapotranspiration occurs at the potential rate as long as the soil
197 moisture is between the field capacity and evapotranspiration limit, below which actual
198 evapotranspiration varies linearly as a fraction of the potential evapotranspiration. The runoff
199 response routine transforms excess water from the soil moisture routine to discharge. It contains
200 two reservoirs, i.e. an upper zone that has two outlets (one linear interflow and one nonlinear near-
201 surface flow) and a lower zone, which has one linear outlet. Finally, the released runoff is filtered
202 by a triangular weighting function (Bergström, 1995; Kollat et al., 2012).

Fig. 2. Structure of HBV model used in this study. Parameters are in red, referring to table 1 for detailed explanation.

In this study, daily potential evapotranspiration is calculated using the Penman-Monteith equation (McMahon et al., 2013). Due to the uneven distribution of gauge stations, the common Thiessen polygon method is used to calculate area-averaged forcing data for each catchment. Model parameters (Table 2) are assumed constant within a catchment and are independently calibrated by

fitting the simulated streamflow against the observed one. Given that periods of calibration and validation are short, the model is probably affected by unreasonable initial conditions. To specify the initial conditions of the catchments, the two-year period is used for model warm-up. Optimization is conducted within MATLAB by calling the embedded genetic algorithm (GA) (Seibert, 2000). By performing selection, crossover, and mutation over a number of generations, the GA is able to derive the optimum set of model parameters given the input and forcing variables (Sreekanth and Datt, 2012). Population size and number of generations are set to 100 and 500, respectively, thus the number of model runs is 50000 for each calibration. The objective function is maximizing the agreement between simulated and observed discharge, which is evaluated by the Kling-Gupta Efficiency (KGE) (Gupta et al., 2009).

Model calibration and validation are carried out at daily scale for four different model setups, i.e. using four different precipitation datasets (IMERG-E, IMERG-F, TRMM 3B42, and gauged dataset). It should be noted that the four setups are independent from each other. All setups use the same model forcing except precipitation and thus, differences between the model efficiency are a result of differences in the precipitation datasets and the corresponding model parameters. The present calibration strategy allows to investigate the utility of satellite datasets for hydrological modeling in ungauged or data-scarce areas, where no ground gauge records are available for model calibration. Model performance is first assessed by KGE. However, considering that KGE can be very different due to the diversity of catchments (Seibert et al., 2018), we also use a skill score (SS) to assess simulations of models forced with satellite precipitation datasets against models forced with gauge-based precipitation. The SS has a range between negative infinity and one. Positive values indicate skill better than the reference, i.e. rain gauge precipitation, while negative values imply simulation worse than the reference. This measure allows to evaluate IMERG-E, IMERG-F, and TRMM 3B42 against their gauge-based counterpart for hydrological simulations. SS is defined by

$$SS = 1 - \frac{\sum(Q_{SimS} - Q_o)^2}{\sum(Q_{SimG} - Q_o)^2} \quad (1)$$

with, Q_{SimS} the simulated discharge when using satellite estimates as forcing, Q_{SimG} the simulated discharge when using gauge-based precipitation as forcing, and Q_o the observed discharge.

If Q_{SimS} has a 10% lower RMSE than Q_{SimG} , then:

$$SS = 1 - \frac{0.9^2 \sum(Q_{SimG} - Q_o)^2}{\sum(Q_{SimG} - Q_o)^2} = 0.19$$

If Q_{SimS} has a 10% higher RMSE than Q_{SimG} , then:

$$SS = 1 - \frac{1.1^2 \sum(Q_{SimG} - Q_o)^2}{\sum(Q_{SimG} - Q_o)^2} = -0.21$$

243

244 **4. Results and analyses**

245 **4.1. Evaluation of satellite estimates against rain gauge observations**

246 *4.1.1. Intercomparison of three products over the entire region*

247 Fig.3 shows the box-whisker plots of rainfall detecting skill scores. IMERG-F outperforms
 248 IMERG-E and TRMM 3B42 with respect to three metrics, i.e. POD, CSI, and HSS, indicating that
 249 IMERG-F yields a higher frequency of successful detection of rainfall events. It should be noted
 250 that IMERG-E outperforms TRMM 3B42 even though the latter is rescaled and calibrated using
 251 GPCC data.

Fig. 3. Box-whisker plots of categorical skill scores over the entire study area. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25% and 75% quartiles, respectively. The whiskers extend to the maximum and minimum, respectively.

Fig. 4 shows the density scatter plots of precipitation derived by the three satellite products at different time scales. At daily scale, IMERG-F and TRMM are almost equally good with respect to all four accuracy metrics. IMERG-E also shows good agreement with gauge-based precipitation but with around -12% underestimation. At monthly scale, IMERG-F and TRMM show similar performance again, but IMERG-F has a slightly higher positive RE. On the contrary, IMERG-E has a 7% negative RE, and a larger spread is obvious when precipitation is below 300 mm/month (Fig. 4). All three products substantially underestimate ($> 20\%$) extreme monthly precipitation (see plots in middle row of Fig. 4). The annual estimates of the three products very well agree with the gauge-based observations with a correlation coefficient over 0.9.

Fig. 4. Scatter plots of daily, monthly, and yearly precipitation from IMERG-F, IMERG-E, and TRMM 3B42 against corresponding gauge-based precipitation. The 1:1 line and its 90% (light grey shade) 95% (dark grey shade) confidence intervals are given.

Clearly, all three products show stronger correlation with rain gauges at longer time scales. Moreover, IMERG-F consistently has a positive RE for all three time scales, while IMERG-E has a negative RE (Fig. 4). TRMM 3B42 has a negative RE at the daily scale, but shows a positive RE at the monthly and annual scale, indicating a season-specific RE. Fig. 5 shows the interannual

variations of the accuracy metrics. Among the four accuracy metrics, the RE shows the most pronounced disagreement among the three products, especially for IMERG-E. Specifically, all months but January have negative biases, which explains the consistent negative bias of IMERG-E at different time scales. Furthermore, a larger disagreement appears in the cold season, i.e. November to February, in terms of RE and CC. TRMM 3B42 shows a large underestimation in the cold season.

Fig. 5. Interannual variations of accuracy metrics of three products. All daily data of each month are pooled together to calculate the metrics.

4.1.2. Analysis over different elevation bands and climate zones

We report performance of the three satellite products over different elevation bands. Fig. 6 shows box-whisker plots of 8 statistical metrics. It is clear that the performances of rainfall detection and accuracy differ markedly over different elevation bands for all products. Overall, all products perform best over low elevations. The correlation coefficient (CC) generally decreases

with increasing elevation, while the categorical metrics show different patterns. Specifically, POD, CSI, and HSS first decrease, and then increase with increasing elevation (Fig. 6), while FAR shows the opposite behavior. Both low and high elevation bands have better performance than middle elevation bands. Note that the metrics cover a wider range (i.e. larger variation) for the elevation band between 1000 and 2000 m, which indicates satellite estimate quality also varies across regions.

Comparing the three products, differences appear over high elevation bands (> 3000 m) and mainly between IMERG-F/-E and TRMM 3B42. TRMM 3B42 has worse detection skills for the highest two elevation bands than its counterparts. Moreover, IMERG-F and TRMM 3B42 tend to overestimate while IMERG-E underestimates precipitation over higher elevation bands (see bias in Fig. 6).

Fig. 6. Box plots of eight metrics for eight elevation bands at grid scale. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25% and 75% quartiles, respectively. The whiskers extend to the maximum and minimum, respectively. A third order polynomial fitting curve is added for each metric.

Fig. 7 shows the performance metrics for different climate zones. It is obvious that all three products perform worst over the arid zone (smallest CC, POD, CSI, and HSS values). All products consistently underestimate over the tropical zone and conversely, overestimate over the arid zone. Nevertheless, the tropical zone shows the best performance in terms of CC and HSS. Note that, TRMM 3B42 has the lowest detection skill and the high frequency of false alarm rate over the arid zone although the mean bias is close to neutral.

Fig. 7. Box-whisker plots of eight metrics for five Köppen-Geiger climate zones.

4.1.3. Spatial variability

For the sake of brevity and clarity, the spatial evaluation will focus on two representative metrics, i.e. HSS and RE. Fig. 8 illustrates the rainfall detection skill of the three products in terms of HSS. Overall, IMERG-F performs much better in the south than in the northwest. According to the Köppen-Geiger climate classification, poorly performing grids (i.e. black dots in Fig. 8a) are mainly located in arid areas, such as Tarim Basin, Qaidam Basin, Tian Shan mountainous areas, and also in temperate areas, i.e., Sichuan Basin. These regions are generally within/close to the elevation band between 1000 and 2000 m (Fig. 8b), over which IMERG-F shows poor detection skill as revealed in the previous section. Comparatively, the HSS of IMERG-E degrades at 74% of the stations, and larger degradations ($-0.2 \sim -0.05$) mainly occur over cold and arid zones (see black dots in Fig. 8c). IMERG-E also shows improvements over some stations; however, this quantity is very small (< 0.05). However, TRMM 3B42 shows a large decrease in HSS compared to IMERG-F. Clearly, this degradations occur over arid, cold, and polar zones (see black dots in Fig. 8e), where elevation is generally higher (Fig. 8f), which agrees with the previous results. TRMM 3B42 has a lower HSS over higher elevation bands (Fig.6). Interestingly, TRMM 3B42 shows large improvements (i.e. crosses and triangles) over the north China plain area (Fig. 8f).

Fig. 8. Detection skill of three products in terms of HSS over different climate zones (left panel) and elevation bands (right panel). Upper row displays HSS of IMERG-F. Middle and lower rows demonstrate the relative change of HSS of IMERG-E and TRMM referring to IMERG-F. Positive value indicates better skill than IMERG-F, and negative for worse skill.

Fig. 9 displays the spatial distribution of over-/underestimation of the three products. Noticeably, the spatial patterns and amplitudes are different among the three products. With regard to IMERG-F, positive biases dominate. Overestimation is pronounced over the Tian Shan mountains, the Qilian mountains, and the Himalayas (Figs. 9a and 9b). IMERG-E has a similar pattern of overestimation, but also shows substantial underestimation over the Tibetan Plateau and the southwestern mountainous areas (Figs. 9c and 9d), which offsets the positive bias and leads to a negative bias over the whole study area. TRMM 3B42 exhibits a very similar pattern compared to IMERG-F but also with some improvements over the north China plain area. However, the underestimation over the Tarim Basin and Junggar Basin is larger (Figs. 9e and 9f).

Fig. 9. Spatial distribution of relative error (RE) showing the accuracy of satellite estimates (IMERG-F, IMERG-E and TRMM 3B42 from top to bottom) over different climate zones (left panel) and elevation bands (right panel).

4.2. Hydrological evaluation

4.2.1. Statistical analysis

Table 3 summarizes the performance metrics of model simulations using the four precipitation inputs. Overall, the models perform relatively well ($KGE > 0.5$) using IMERG-F, IMERG-E, and gauge-based precipitation during calibration period. In terms of SS, IMERG-F is equally good as the gauge-based precipitation while IMERG-E is also comparable. However, TRMM 3B42 is worse with a negative SS value (-0.22), which is consistent with a lower KGE value (0.42). During the validation period, all performance statistics are not satisfactory ($KGE < 0.3$). Comparatively, IMERG-F and IMERG-E are better than gauge-based precipitation as indicated by positive SS values and larger KGE values (Table 3), while TRMM 3B42 performs worse than gauge-based precipitation.

Regarding climate conditions, all four datasets provide good performance in the temperate climate zone and poor performance in the arid zone. Compared with gauge-based precipitation, IMERG-F performs better in all zones except the cold zone. This is more pronounced during the validation period (see positive SS in Table 3). IMERG-E also shows a promising performance except in the polar zone. In contrast, TRMM 3B42 consistently performs worse than the gauge-based precipitation except in the tropical zone in terms of KGE and SS. Interestingly, the RE distribution does not agree with the other metrics. The better values for RE (around 0) point into a different direction than for CC, KGE and SS (towards 1). For example, although RE is relatively small for TRMM 3B42, CC, KGE, and SS are relatively low.

Table 3. Performance metrics of model simulation using four precipitation inputs (median value is used for each metrics). Statistics are also reported by five climate zones.

Climate zone	Number of catchments	Dataset	Calibration				Validation			
			RE (%)	CC	KGE	SS	RE (%)	CC	KGE	SS
Tropical (A)	2	IMERG-E	-1.8	0.77	0.87	0.56	-51.9	0.14	-0.20	0.24
		IMERG-F	45.5	0.64	0.50	0.23	47.2	0.30	0.14	0.28
		3B42	11.0	0.68	0.77	0.40	3.9	0.28	0.35	0.50
		Gauge	35.9	0.55	0.54	0	58.6	0.26	0.07	0
Arid (B)	10	IMERG-E	-0.8	0.55	0.48	0.27	-13.6	0.11	0.07	0.83

Temperate (C)	231	IMERG-F	-2.8	0.43	0.35	0.18	-46.8	0.12	-0.02	0.78
		TRMM 3B42	-8.2	0.47	0.16	-0.13	70.9	0.03	-0.71	-14.0
		Gauge	10.5	0.45	0.49	0	>300	0.11	-318	0
		IMERG-E	-0.5	0.45	0.60	-0.06	-16.4	0.31	0.21	0.07
		IMERG-F	3.1	0.46	0.62	0.03	2.5	0.39	0.31	0.12
		TRMM 3B42	-1.2	0.31	0.50	-0.17	6.4	0.21	0.04	-0.06
		Gauge	1.4	0.44	0.62	0	8.5	0.27	0.16	0
		IMERG-E	5.9	0.41	0.57	0.0	-19.3	0.12	-0.18	0.28
		IMERG-F	7.3	0.42	0.54	-0.10	-7.5	0.18	-0.24	0.15
Cold (D)	55	TRMM 3B42	6.5	0.12	0.13	-0.49	168	0.05	-3.70	-0.59
		Gauge	3.1	0.48	0.64	0	67.6	0.15	-0.69	0
		IMERG-E	-26.2	0.70	0.64	-0.49	-23.5	0.63	0.42	-0.51
		IMERG-F	-13.3	0.73	0.76	0.23	-11.5	0.74	0.49	0.08
Polar (E)	2	TRMM 3B42	-41.8	0.27	0.19	-2.13	199	0.11	-28.2	-930
		Gauge	-18.6	0.60	0.66	0	-31.0	0.44	0.43	0
		IMERG-E	0.1	0.45	0.59	-0.05	-16.9	0.26	0.16	0.10
		IMERG-F	3.3	0.45	0.60	-0.01	1.2	0.34	0.23	0.13
All	300	TRMM 3B42	-0.35	0.26	0.42	-0.22	9.5	0.14	-0.14	-0.09
		Gauge	1.8	0.45	0.61	0	14.9	0.23	0.08	0

367

368 Fig. 10 displays the model efficiency in terms of SS as a function of catchment size. Regarding
369 IMERG-F, no noticeable difference between the different size categories is revealed, which
370 indicates that catchment area does not dominate the catchment hydrological response at moderate
371 scales. However, catchments with areas between 4000 and 5000 km² perform slightly better
372 (narrower range of SS, see first column in Fig. 10) probably because the average satellite
373 precipitation in large catchment is more representative. Moreover, some catchments provide
374 positive SS values, which indicates that IMERG-F outperforms gauge-based precipitation in terms
375 of its impact on the hydrological simulations. This can be explained by the low density of gauge
376 networks. IMERG-E shows a similar pattern as IMERG-F except for some extreme values.
377 Comparatively, the distribution of SS is very wide for TRMM 3B42. Similarly, there is no clear
378 differences among these five size categories.

Fig. 10. Histogram of model performance metrics (SS) using IMERG-F, IMERG-E, and TRMM as input during calibration period. Each row from top to bottom shows results of catchments with area of [< 1000], [$1000, 2000$], [$2000, 3000$], [$3000, 4000$], [$4000, 5000$] and [$0, 5000$] km², respectively.

Similarly, we investigate the hydrological performance as a function of catchment elevation. As shown in Fig. 11, no clear pattern can be discerned. Comparatively, the highest elevation group exhibits a slightly poorer performance revealed by a smaller SS value. However, the

387 difference becomes larger for TRMM 3B42, and the third group ([874, 1350] m) shows a better
 388 performance.

389

390 Fig. 11. Histogram of model performance metrics (SS) using IMERG-F, IMERG-E, and TRMM
 391 as input during calibration period. Each row from top to bottom shows results of catchments with
 392 average elevation of [< 375], [375, 874], [874, 1350], [1350, 4261], and [0, 4261] m,
 393 respectively.

394 4.2.2. Spatial variability

395 As shown in Fig. 12, the benchmark model (i.e. forced with gauge-based precipitation)
 396 performs reasonably well ($KGE > 0.6$) in most catchments. Nevertheless, 22% of all catchments

have a low KGE value (< 0.4); those catchments are mainly distributed in Weihe river basin (middle part of the Yellow River) and the Sichuan basin. This can be explained by the orographic effects and/or relatively coarse distribution of rainfall gauges, which do not well represent catchment precipitation. On the contrary, during validation period, the performance is not satisfactory (< 0.4) except for 25% of the catchments, which are mainly located in the south of the Yangtze River (Fig. 12b). This is most probably attributed to intra-annual variation of streamflow.

Compared with the benchmark model, the model forced with IMERG-F performs better over those catchments with smaller KGE value provided by benchmark model (see blue-coded one in Fig. 12c and yellow-/orange coded ones in Fig. 12a). For catchments located in the central part, models forced with IMERG-F perform not as well as the benchmark. It should be noted that, models perform satisfactorily over these catchments in terms of the KGE value as shown in Table 3. Here we highlight the difference between model forced with satellite precipitation and those forced with gauge-based precipitation. During the calibration period, model performance improves over those with negative KGE values (red-coded in Fig. 12b) provided by benchmark model.

Regarding IMERG-E, the hydrological performance is comparable to IMERG-F. But the spatial distribution of model performance using TRMM 3B42 is very different especially in calibration period (Fig. 12g). Performance improves markedly in the central part (around 110°E), where IMERG-E and IMERG-F do not outperform gauge-based precipitation.

Fig. 12. Spatial distribution of model performance metrics. Left and right panels show results of calibration and validation, respectively. Top row shows KGE value of model simulation using gauge-based precipitation. Bottom three rows show SS value of model simulation.

5. Discussion

In general, IMERG-E does not show remarkably large difference from IMERG-F when compared with gauge-based precipitation. Moreover, both IMERG-E and IMERG-F show an improvement in detecting rainfall events compared to TRMM 3B42. However, it should be noted that this evaluation is conducted at pixel-to-point scale. The spatial resolution of the IMERG products and TRMM 3B42 are 0.1° and 0.25° , respectively. That means, we are comparing the mean precipitation of a large area ($> 100 \text{ km}^2$) with the one at a point, which is especially problematic in heterogeneous terrain with variation in topography, climate conditions, etc. In this context, it makes sense that IMERG is better than TRMM 3B42 due to its higher spatial resolution. This could probably be the reason why extreme monthly precipitation is underestimated as shown in Figure 4. Moreover, satellite precipitation products typically perform poorly in mountainous areas as reported by previous studies (Dinku et al., 2008; Maggioni et al., 2016; Xu et al., 2017). Our results confirm that satellite products perform poorly over high elevation zones and complex topographic regions, such as the Tarim basin, Tian Shan mountainous areas, Qinling mountainous areas, Sichuan basin, etc. This issue is, however, more pronounced for TRMM 3B42.

We assessed IMERG as a forcing for a rainfall-runoff model. As presented in Table 3, the KGE between model simulations and gauged streamflow is generally satisfactory, with median values of 0.60 during the calibration period. However, low KGE values (median value of 0.23) were observed during the validation period. This can be attributed to the short calibration period, which may not be representative for long-term catchment conditions. Calibration parameters tend to compensate for other factors, such as model structure, forcing data, etc. and therefore, they may change for different calibration periods, especially for short periods (Merz et al., 2011). These calibrated parameters may produce large errors in simulations for other climate conditions. For example, Fig. 13 shows the simulations during the calibration and validation periods. As we can see, the reliability of the ensemble simulation (Latin Hypercube Sampling method) is low in the validation period; a large fraction of the observations fall outside the confidence interval of the simulation. It is also

445 evident that the GA successfully found the optimum parameter set. Therefore, the poor performance
 446 in the validation period may be due to differences in hydrological conditions between the
 447 calibration and validation periods.

448
 449 Fig. 13. Model simulations during the calibration and validation periods, demonstrating the
 450 variations of hydrological conditions during this two-year period.

451 No pronounced difference in performance was found in terms of catchment area and elevation,
 452 although precipitation accuracy varies among different elevation zones. This indicates that the

uncertainty of precipitation did not necessarily translate into bad runoff simulations. Overall, the model performs best in temperate climate zones.

Moreover, model simulations forced with IMERG-F and IMERG-E estimates show similar performance as those forced with gauge-based precipitation. IMERG-E slightly underestimates precipitation, but the hydrological simulation is acceptable. On the contrary, although TRMM 3B42 shows a comparable agreement with gauge-based precipitation, the performance in hydrological simulations is relatively poor. One possible explanation is the spatial variability, which needs further investigation with high density gauge networks. Moreover, the relation between the quality of the precipitation product and the hydrological performance is not straightforward as reported by previous studies (Camici et al., 2018; Qi et al., 2016).

Overall, the transformation of IMERG estimates into runoff is satisfactory across China. Given that the density of the gauging station network is quite sparse in China, IMERG has potential as an alternative to gauge-based precipitation. The findings of this study show the hydrological performance at medium catchment scale. It can be directly used in hydrological modeling without bias correction in most of the catchments.

While our results provide insights into the accuracy of IMERG estimates and the suitability of IMERG Early and Final runs for hydrological modeling, it should be noted that there are some assumptions and limitations. The hydrological performance is solely based on the lumped HBV model. Hence, the results would be different if a different rainfall-runoff model was used. The short period is another limitation of this study. However, we believe a timely assessment is most important for further product improvement and future research.

6. Summary and conclusions

IMERG products inter-calibrate, merge, and interpolate a variety of satellite microwave precipitation estimates, microwave-calibrated infrared satellite estimates, and precipitation gauge analyses at high temporal (half-hour) and spatial resolution ($0.1^\circ \times 0.1^\circ$). This study provides an early and timely evaluation and hydrological application of the early (IMERG-E) and final run (IMERG-F) products of IMERG during two-year period (2016 - 2017) over China.

We have assessed daily IMERG precipitation estimates using rain gauge records. For the evaluation, we computed both continuous and categorical metrics to evaluate data accuracy and the ability to identify rainfall events. We put special emphasis on climate condition (Köppen-Geiger climate classification) and topographic condition (elevation). The main findings are:

- 1) In terms of the Heidke skill score (HSS), the performance of the three evaluated products in descending order is IMERG-F, IMERG-E, TRMM 3B42. Nevertheless, IMERG-E underestimates by around 12%.
- 2) Statistical metrics are very diverse across different climate zones. All products consistently underestimate over the tropical zone and conversely, overestimate over the arid zones.
- 3) IMERG-E and IMERG-F show a slight dependence on topography and have poor performance over high elevation zones. TRMM 3B42 performs much worse over high elevation zones.

Care should be taken when using IMERG products, and especially TRMM 3B42, in some regions such as the Tarim basin, Tian Shan mountainous areas, Qinling mountainous areas, Sichuan basin, etc. Bias correction may be required to reduce uncertainty.

In order to understand how good IMERG-E and IMERG-F are for hydrological modeling, we conducted rainfall-runoff modeling over 300 catchments across China using a typical lumped model. The specific conclusions from our analysis are:

- 1) Overall, models forced with IMERG-E and IMERG-F perform as well as those forced with gauge-based precipitation in most cases, and much better than those forced with TRMM 3B42.
- 2) IMERG-E estimates can be directly used to inform operational rainfall-runoff modeling in most catchments with competitive performance.
- 3) However, region-specific discrepancies remain, e.g. model performance is much better in humid regions.

In summary, IMERG gives acceptable performance in precipitation estimates and transformation of precipitation into runoff across China. However, the performance varies across climate regions and topography. Care should be taken when forcing hydrological models with IMERG in the arid regions.

Acknowledgements

This work is partly funded by the Ministry of Foreign Affairs of Denmark and administered by Danida Fellowship Centre (File number: 18-M01-DTU). We would like to thank Hannes Müller-Thomy and two anonymous reviewers for their suggestions and comments that greatly contributed to improving the final version of this paper. Thanks also go to the Associate Editor for efficiently managing the peer review process. We also acknowledge the China Meteorological Administration for sharing the ground-based precipitation data. The IMERG data are provided by the NASA/Goddard Space Flight Center's PMM Science Team and Precipitation Processing System (PPS), which develop and compute the IMERG data as a contribution to GPM, and archived at the NASA Goddard Earth Science Data and Information System Center (GES DISC).

References

Ashouri, H., Nguyen, P., Thorstensen, A., Hsu, K., Sorooshian, S., Braithwaite, D., 2016.

523 Assessing the Efficacy of High-Resolution Satellite-Based PERSIANN-CDR Precipitation
 524 Product in Simulating Streamflow. *J. Hydrometeorol.* 17, 2061–2076. doi:10.1175/JHM-D-
 525 15-0192.1

526 Beck, H.E., van Dijk, A.I.J.M., de Roo, A., Dutra, E., Fink, G., Orth, R., Schellekens, J., 2016.
 527 Global evaluation of runoff from ten state-of-the-art hydrological models. *Hydrol. Earth*
 528 *Syst. Sci. Discuss.* 1–33. doi:10.5194/hess-2016-124

529 Beck, H.E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A.I.J.M., Weedon, G.P., Brocca,
 530 L., Pappenberger, F., Huffman, G.J., Wood, E.F., 2017. Global-scale evaluation of 22
 531 precipitation datasets using gauge observations and hydrological modeling. *Hydrol. Earth*
 532 *Syst. Sci.* 21, 6201–6217. doi:10.5194/hess-21-6201-2017

533 Bergström, S., 1995. The HBV model, in: *Computer Models of Watershed Hydrology*. Water
 534 Resources Publications, Highlands Ranch, Colorado, USA, pp. 443–476.

535 Brocca, L., Moramarco, T., Melone, F., Wagner, W., 2013. A new method for rainfall estimation
 536 through soil moisture observations. *Geophys. Res. Lett.* 40, 853–858. doi:10.1002/grl.50173

537 Camici, S., Ciabatta, L., Massari, C., Brocca, L., 2018. How reliable are satellite precipitation
 538 estimates for driving hydrological models: A verification study over the Mediterranean area.
 539 *J. Hydrol.* 563, 950–961. doi:10.1016/j.jhydrol.2018.06.067

540 Ciabatta, L., Brocca, L., Massari, C., Moramarco, T., Gabellani, S., Puca, S., Wagner, W., 2016.
 541 Rainfall-runoff modelling by using SM2RAIN-derived and state-of-the-art satellite rainfall
 542 products over Italy. *Int. J. Appl. Earth Obs. Geoinf.* 48, 163–173.
 543 doi:10.1016/j.jag.2015.10.004

544 Dinku, T., Chidzambwa, S., Ceccato, P., Connor, S.J., Ropelewski, C.F., 2008. Validation of
 545 high-resolution satellite rainfall products over complex terrain. *Int. J. Remote Sens.* 29,
 546 4097–4110. doi:10.1080/01431160701772526

547 Guo, H., Chen, S., Bao, A., Behrangi, A., Hong, Y., Ndayisaba, F., Hu, J., Stepanian, P.M., 2016.
 548 Early assessment of Integrated Multi-satellite Retrievals for Global Precipitation
 549 Measurement over China. *Atmos. Res.* 176–177, 121–133.
 550 doi:10.1016/j.atmosres.2016.02.020

551 Gupta, H. V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared
 552 error and NSE performance criteria: Implications for improving hydrological modelling. *J.*
 553 *Hydrol.* 377, 80–91. doi:10.1016/j.jhydrol.2009.08.003

554 He, Z., Yang, L., Tian, F., Ni, G., Hou, A., Lu, H., 2017. Intercomparisons of Rainfall Estimates
 555 from TRMM and GPM Multisatellite Products over the Upper Mekong River Basin. *J.*
 556 *Hydrometeorol.* 18, 413–430. doi:10.1175/JHM-D-16-0198.1

557 Huffman, G.J., Bolvin, D.T., Braithwaite, D., Hsu, K., Joyce, R.J., Kidd, C., Nelkin, E.J., Xie, P.,
 558 2015. Algorithm Thoretical Basis Document (ATBD).

559 Huffman, G.J., Bolvin, D.T., Nelkin, E.J., 2017. Integrated Multi - satellitE Retrievals for GPM
 560 (IMERG) Technical Documentation.

561 Huffman, G.J., Bolvin, D.T., Nelkin, E.J., Wolff, D.B., Adler, R.F., Gu, G., Hong, Y., Bowman,
 562 K.P., Stocker, E.F., 2007. The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-
 563 Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales. *J.*
 564 *Hydrometeorol.* 8, 38–55. doi:10.1175/JHM560.1

565 Joyce, R.J., Janowiak, J.E., Arkin, P.A., Xie, P., 2004. CMORPH: A Method that Produces
566 Global Precipitation Estimates from Passive Microwave and Infrared Data at High Spatial
567 and Temporal Resolution. *J. Hydrometeorol.* 5, 487–503. doi:10.1175/1525-
568 7541(2004)005<0487:CAMTPG>2.0.CO;2

569 Kollat, J.B., Reed, P.M., Wagener, T., 2012. When are multiobjective calibration trade-offs in
570 hydrologic models meaningful? *Water Resour. Res.* 48, 1–19. doi:10.1029/2011WR011534

571 Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., 2006. World Map of the Köppen-Geiger
572 climate classification updated. *Meteorol. Zeitschrift* 15, 259–263. doi:10.1127/0941-
573 2948/2006/0130

574 Lehner, B., Grill, G., 2013. Global river hydrography and network routing: baseline data and new
575 approaches to study the world's large river systems. *Hydrol. Process.* 27, 2171–2186.
576 doi:10.1002/hyp.9740

577 Li, N., Tang, G., Zhao, P., Hong, Y., Gou, Y., Yang, K., 2017. Statistical assessment and
578 hydrological utility of the latest multi-satellite precipitation analysis IMERG in Ganjiang
579 River basin. *Atmos. Res.* 183, 212–223. doi:10.1016/j.atmosres.2016.07.020

580 Liu, Z., 2016. Comparison of Integrated Multisatellite Retrievals for GPM (IMERG) and TRMM
581 Multisatellite Precipitation Analysis (TMPA) Monthly Precipitation Products: Initial
582 Results. *J. Hydrometeorol.* 17, 777–790. doi:10.1175/JHM-D-15-0068.1

583 Maggioni, V., Meyers, P.C., Robinson, M.D., 2016. A Review of Merged High-Resolution
584 Satellite Precipitation Product Accuracy during the Tropical Rainfall Measuring Mission
585 (TRMM) Era. *J. Hydrometeorol.* 17, 1101–1117. doi:10.1175/JHM-D-15-0190.1

586 McMahon, T.A., Peel, M.C., Lowe, L., Srikanthan, R., McVicar, T.R., 2013. Estimating actual,
587 potential, reference crop and pan evaporation using standard meteorological data: A
588 pragmatic synthesis. *Hydrol. Earth Syst. Sci.* 17, 1331–1363. doi:10.5194/hess-17-1331-
589 2013

590 Mei, Y., Nikolopoulos, E.I., Anagnostou, E.N., Borga, M., 2016. Evaluating Satellite
591 Precipitation Error Propagation in Runoff Simulations of Mountainous Basins. *J.*
592 *Hydrometeorol.* 17, 1407–1423. doi:10.1175/JHM-D-15-0081.1

593 Merz, R., Parajka, J., Blöschl, G., 2011. Time stability of catchment model parameters:
594 Implications for climate impact analyses. *Water Resour. Res.* 47, 1–17.
595 doi:10.1029/2010WR009505

596 Prakash, S., Mitra, A.K., Pai, D.S., AghaKouchak, A., 2016. From TRMM to GPM: How well
597 can heavy rainfall be detected from space? *Adv. Water Resour.* 88, 1–7.
598 doi:10.1016/j.advwatres.2015.11.008

599 Qi, W., Zhang, C., Fu, G., Sweetapple, C., Zhou, H., 2016. Evaluation of global fine-resolution
600 precipitation products and their uncertainty quantification in ensemble discharge
601 simulations. *Hydrol. Earth Syst. Sci.* 20, 903–920. doi:10.5194/hess-20-903-2016

602 Seibert, J., 2000. Multi-criteria calibration of a conceptual runoff model using a genetic
603 algorithm. *Hydrol. Earth Syst. Sci.* 4, 215–224. doi:10.5194/hess-4-215-2000

604 Seibert, J., Vis, M.J.P., Lewis, E., van Meerveld, H.J., 2018. Upper and lower benchmarks in
605 hydrological modelling. *Hydrol. Process.* doi:10.1002/hyp.11476

606 Sorooshian, S., Hsu, K.-L., Gao, X., Gupta, H. V., Imam, B., Braithwaite, D., 2000. Evaluation of

- PERSIANN System Satellite–Based Estimates of Tropical Rainfall. *Bull. Am. Meteorol. Soc.* 81, 2035–2046. doi:10.1175/1520-0477(2000)081<2035:EOPSSE>2.3.CO;2
- Sreekanth, J., Datt, B., 2012. Genetic Programming: Efficient Modeling Tool in Hydrology and Groundwater Management, in: *Genetic Programming - New Approaches and Successful Applications*. InTech, pp. 32–35. doi:10.5772/52102
- Tang, G., Zeng, Z., Long, D., Guo, X., Yong, B., Zhang, W., Hong, Y., 2016. Statistical and Hydrological Comparisons between TRMM and GPM Level-3 Products over a Midlatitude Basin: Is Day-1 IMERG a Good Successor for TMPA 3B42V7? *J. Hydrometeorol.* 17, 121–137. doi:10.1175/JHM-D-15-0059.1
- Tian, Y., Peters-Lidard, C.D., Choudhury, B.J., Garcia, M., 2007. Multitemporal Analysis of TRMM-Based Satellite Precipitation Products for Land Data Assimilation Applications. *J. Hydrometeorol.* 8, 1165–1183. doi:10.1175/2007JHM859.1
- Wang, C., Tang, G., Han, Z., Guo, X., Hong, Y., 2018. Global intercomparison and regional evaluation of GPM IMERG Version-03, Version-04 and its latest Version-05 precipitation products: Similarity, difference and improvements. *J. Hydrol.* 564, 342–356. doi:10.1016/j.jhydrol.2018.06.064
- Wang, Z., Zhong, R., Lai, C., Chen, J., 2017. Evaluation of the GPM IMERG satellite-based precipitation products and the hydrological utility. *Atmos. Res.* 196, 151–163. doi:10.1016/j.atmosres.2017.06.020
- Xiao, M., Zhang, Q., Singh, V.P., Chen, X., 2013. Regionalization-based spatiotemporal variations of precipitation regimes across China. *Theor. Appl. Climatol.* 114, 203–212. doi:10.1007/s00704-013-0832-1
- Xu, R., Tian, F., Yang, L., Hu, H., Lu, H., Hou, A., 2017. Ground validation of GPM IMERG and TRMM 3B42V7 rainfall products over southern Tibetan Plateau based on a high-density rain gauge network. *J. Geophys. Res. Atmos.* 122, 910–924. doi:10.1002/2016JD025418
- Zambrano-Bigiarini, M., Nauditt, A., Birkel, C., Verbist, K., Ribbe, L., 2017. Temporal and spatial evaluation of satellite-based rainfall estimates across the complex topographical and climatic gradients of Chile. *Hydrol. Earth Syst. Sci.* 21, 1295–1320. doi:10.5194/hess-21-1295-2017
- Zubieta, R., Getirana, A., Espinoza, J.C., Lavado-Casimiro, W., Aragon, L., 2017. Hydrological modeling of the Peruvian–Ecuadorian Amazon Basin using GPM-IMERG satellite-based precipitation dataset. *Hydrol. Earth Syst. Sci.* 21, 3543–3555. doi:10.5194/hess-21-3543-2017