Blending multi-satellite, atmospheric reanalysis and gauge precipitation products to facilitate hydrological modelling

Jiabo Yin[1,2], Shenglian Guo[1]*, Lei Gu[3], Ziyue Zeng[4], Dedi Liu[1,2], Jie Chen[1,2], Youjiang Shen[1], Chong-Yu Xu[5]

[1] State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan 430072, China

[2] Hubei Provincial Key Lab of Water System Science for Sponge City Construction, Wuhan University, Wuhan 430072, China

[3] School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

[4] Changjiang River Scientific Research Institute, Wuhan 430015, China

[5] Department of Geosciences, University of Oslo, P.O. Box 1047 Blindern, N-0316 Oslo, Norway

Abstract: Satellite-retrieved and atmospheric reanalysis precipitation can bridge the spatiotemporal gaps of in-situ gauging networks, but estimation biases can limit their reliable applications in hydrological monitoring and modelling. To correct precipitation occurrence and intensity simultaneously, this study develops a three-stage blending approach to integrate three multi-satellite precipitation datasets (IMERG Final, TMPA 3B42V7 and PERSIANN-CDR), the ERA5 atmospheric reanalysis product and a gauge dataset within a dynamic framework. Firstly, the systematic biases of the four members were individually corrected by combining the

* Correspondence author: slguo@whu.edu.cn (S. Guo)
E-mail address: jboyn@whu.edu.cn (J. Yin); shisan@hust.edu.cn (L. Gu); zengzy@mail.crsri.cn (Z. Zeng); dediliu@whu.edu.cn (D. Liu); jiechen@whu.edu.cn (J. Chen); yjshen@whu.edu.cn (Y. Shen); c.y.xu@geo.uio.no (C.-Y. Xu)

local intensity scaling and ratio bias correction methods. Then, the "state weights" used for determining wet/dry events were optimized by evaluating a score function of the four bias-corrected members. Thirdly, the "intensity weights" were optimized using the cuckoo search (CS) algorithm and the Bayesian Model Averaging (BMA) method, respectively. The three-stage blending approach produced dynamic weights varying both spatially and temporally, and the performance was thoroughly evaluated over mainland China. Results show that the three-stage dynamic scheme performs better than individual datasets and two-stage blending methods in terms of all eight statistical metrics, and the CS algorithm outperforms the BMA method in the third stage. By randomly sampling validation sites using K-fold experiments, the developed algorithm also demonstrates a superior performance in ungauged regions. After interpolating and normalizing blending parameters of all gauges to entire domain using ordinary kriging, a new blended precipitation dataset with a daily 0.25° scale was produced. Four hydrological models are forced by blended and primary precipitations in 238 catchments over China, further confirming that the developed approach can facilitate hydrological modelling demonstrated by improving the Kling-Gupta efficiency of simulated streamflow by 12-35%.

Keywords: Satellite precipitation; Atmospheric reanalysis; Bias correction; Data fusion; Hydrological modelling; China

1. Introduction

As precipitation is a key element in global water cycle and a fundamental forcing in hydrological processes, its accurate observation is of immense importance for decision-making and planning across diverse fields such as hydrology, meteorology, climate and agriculture (Amjad et al., 2020; Yang et al., 2020). The in-situ gauge instrumentation and radar networks might document precipitation with a high accuracy; however, these networks are usually unevenly and sparsely distributed, failing to capture the spatiotemporal heterogeneity of precipitation patterns (Tang et al., 2016), particularly in economically underdeveloped regions such as the western China. With rapid advances in remote sensing technologies and climate system modelling in recent decades, spaceborne sensors and state-of-the-art numerical weather models have produced vast precipitation datasets with a near-global coverage and an unprecedented spatiotemporal resolution (Sunilkumar et al., 2016; Prakash et al., 2018). For example, the Integrated Multi-satellitE Retrievals for GPM (IMERG) product is now available at 0.1° spatial and 30-minute temporal resolutions, and is anticipated to play a growing role in hydrological and meteorological monitoring (Beria et al., 2017; Massari et al., 2020). One of the latest global atmospheric reanalysis products, ERA5 provides hourly estimates for a large number of atmospheric, land and oceanic climate variables, and exhibits substantial improvements comparing to its predecessor in many regions of the globe (Graham et al., 2019; Tang et al., 2020; Tarek et al., 2020).

Although these satellite-based and atmospheric reanalysis precipitation estimates have been widely used in a broad range of applications, their performances are highly constrained by errors and remain an area for further improvements (Bharti and Singh, 2015; Wright et al., 2017; Luo et al., 2019). Ensemble-based approaches, enabling to synthesize multi-source information, generally produce merging precipitations with a better performance than individual members (Beck et al., 2019). While these approaches differ in statistical tools and data sources, they could be classified into two types in terms of one stage or more stages used in data processing. The first type of approach directly combines different multi-source precipitations by optimizing weights based on gauging observations to reduce estimate errors. For example, Massari (2019) developed a Bayesian inversion approach to integrate the TRMM 3B42RT product with the soil moisture-based rainfall dataset SM2RAIN-CCI, and verified its performance in Italy. Among the one-stage blending type, the inverse error variance weighting (Huffman et al., 1997), simple model averaging (SMA) method (Raftery et al., 2005), one-outlier removed method (Shen et al., 2014), Bayesian model averaging (Chao et al., 2018), paired sample t-test and principal component analysis method (Rahman et al., 2018), as well as the machine learning techniques (Bhuiyan et al., 2019) can also achieve better skills than the individual members. A second type of blending approach is emerging following a general concept, i.e., eliminating biases in individual datasets and then merging the bias-adjusted estimates with point-wise gauge observations. For instance, Li et al. (2015) removed the systematic errors in the CMORPH satellite estimates using a bias correction procedure, and then employed a two-dimensional variational analysis scheme to combine the adjusted satellite data with gauging observations.

Due to the spatiotemporal heterogeneity of precipitation, the merging weights of multi-source data might dynamically vary in space and time. To account for such dynamic property of the weights used in integration schemes, Ma et al. (2018a) proposed a dynamic Bayesian model averaging (BMA) algorithm for blending satellite precipitation data from TMPA 3B42RT, 3B42V7, CMORPH and PERSIANN-CDR, and the validation results in the Tibetan plateau verified that the dynamic BMA algorithm outperformed traditional ensemble methods. Rahman et al. (2020) developed a dynamic clustered BMA method and verified that it can accommodate the spatiotemporal differences of diverse satellite products, thus improving precipitation estimation quality even in regions with complex climate and topography patterns. However, these efforts might be insufficient for blending multi-source precipitation in meteorological and hydrological applications. Although the errors in individual members have received substantial attention in some blending techniques (e.g., Li et al., 2015; Beck et al., 2019), some recent dynamic blending approaches neglect this issue (Ma et al., 2018a, b; Rahman et al., 2020). More importantly, satellite/reanalysis products are plagued by problems associated with a lack of precipitation detection, false detection and bias (Maggioni et al., 2016). Existing merging studies usually focused on correcting precipitation magnitude, but failed to minimize bias and eliminate the lack of detection and false alarms simultaneously. For example, the dynamic blending method is only applicable in cases where a rain event hits (Ma et al., 2018a, b; Rahman et al., 2020), which may result in many un-avoided problems. It would be difficult to define a ‘true’ precipitation event if no preprocessing is used to improve the precipitation detection capacity of individual products. Moreover, neglecting the capacity of detecting a wet/dry event not only results in over-estimated or under-estimated rainy days, but also affects the quantitative estimation of precipitation intensity (Tobin and Bennett, 2010), thus inhibiting potential applications in hydrological monitoring and modelling.

To address such concerns, this study developed a three-stage approach to integrate three multi-satellite precipitation products (IMERG Final, TMPA 3B42V7 and PERSIANN-CDR), the ERA5 atmospheric reanalysis product and a gauge dataset. To accommodate the spatiotemporal variations of different members' performances, this scheme was implemented within a dynamic framework to produce dynamic weights varying both spatially and temporally. We utilized a range of evaluation metrics to ascertain the estimation skills in capturing precipitation of the developed method, and also evaluated their potential utility in ungauged regions by randomly sampling validation sites by using K-fold experiments. The dynamic parameters were interpolated and normalized with ordinary kriging (OK) approach, and a new blended precipitation dataset over mainland China with a daily and 0.25° spatiotemporal resolution was produced. We finally evaluated the hydrological performance of our dataset and four individuals by driving four hydrological models over 238 catchments varying in size and climate.

2. Study area and data

2.1. Study area

In meteorological and hydrological assessments, mainland China is usually divided into eight sub-regions (Fig. 1) based on monsoon climate characteristics, administrative divisions, topography, water resources and agricultural geographical distribution (Shi and Xu, 2007; Li et al., 2015). This division is adopted in this study to better illustrate the regional evaluation statistics. Precipitation over southeastern China is mainly dominated by organized deep convection, while northeastern China is more associated with large-scale synoptic weather systems (Chen et al., 2009). In western inland regions, precipitation is usually governed by clouds with shallow depth and limited atmospheric moisture (Shi and Xu, 2007); while in Tibetan plateau and the semiarid northwestern China, precipitation is highly determined by scattered cloud systems (Tong et al., 2013; Li et al., 2015). To evaluate the feasibility of using satellite/reanalysis products and their merging estimates in hydrological simulation, 238 catchments across climatically and topographically diverse regions are selected (Fig. 1), with their watershed boundaries delineated by the Hydro-BASINS product of the HydroSHEDS database (Lehner and Grill, 2013).

[Insert Fig. 1 about here]

2.2. In-situ observation datasets

A ground network of 838 in-situ gauges over mainland China is used as reference, which provides daily precipitation observations and is maintained by the China Meteorological Administration. Most of these stations are densely grouped over eastern China, while gauge stations are much sparser over western China (Fig. 1). Daily air temperature data (including maximum, minimum and average temperature) at a 0.5°×0.5° resolution is obtained from the China Climatic Data Service Center (http://data.cma.cn/en). This gridded dataset was produced based on 2472 in-situ observation gauge stations across China using a spline interpolation method and a

GTOPO30 data sampling algorithm, and is usually considered as an observation reference (Zhang et al., 2015). The daily streamflow data of 238 catchments are obtained from the nine water resources management agencies affiliated to the Ministry of Water Resources of China (http://xxfb.mwr.cn/). These different datasets spanning 2004-2018 are used in this study for two purposes: evaluating satellite precipitation estimates and forcing hydrological models.

2.3 Satellite-based and reanalysis precipitation products

Three multi-satellite precipitation datasets retrieved by integrating infrared and passive microwave sensors were used in this study. The GPM Core Observatory carries the first space-borne Ku/Ka-band dual-frequency radar and a multi-channel microwave imager, thus improving the ability in monitoring both light and solid precipitations (Zubieta et al., 2017). Since the first release of IMERG products in 2015, it has undergone many improvements and the latest version V06B has been reprocessed retrospectively to include TRMM-era data from June 2000 afterwards (Huffman et al., 2019). Owing to the infusion of the Global Precipitation Climatology Centre (GPCC) rain gauge data, the IMERG Final run provides more accurate estimates and is therefore adopted in this study. In addition, the TMPA product 3B42V7 was used, which was also corrected with gauge data and available at 0.25° spatial and 3-hourly temporal resolutions covering 50°N to 50°S for the period of 1998-2019 (Huffman et al., 2010). Moreover, a long-term global precipitation dataset, PERSIANN-CDR, is also used in this study. It is generated from the PERSIANN algorithm using GridSat-B1 infrared data and adjusted by data from the Global

Precipitation Climatology Project (Ashouri et al., 2015). The PERSIANN-CDR dataset provides daily precipitation estimates at 0.25° spatial resolution at a near-global (60°S-60°N) coverage over the period from 1983 to present.

A global atmospheric reanalysis precipitation dataset developed by ECMWF, the ERA5, is used as a final precipitation product in this study. ERA5 provides real-time global hourly precipitation records from 1979 to present on 137 vertical levels from the surface up to 0.1 hPa (Nogueira 2020). ERA5 data are produced by combining model simulations and observations using the laws of physics, which are based on data assimilation by the Integrated Forecasting System (IFS Cy31r2). This assimilation system includes a four-dimensional variational (4D-Var) analysis method and considers the exact timing of observations and model evolution within the assimilation window, enabling to estimate biases between observations and to sift good-quality data from poor data (Nogueira, 2020). The hourly output resolution available at 0.25°×0.25° is an improvement with respect to its predecessor ERA-Interim, and thus provides a more sophisticated simulation of weather processes. All the sub-daily satellite/reanalysis data covering 2004-2018 are aggregated into a daily scale, and the IMERG Final data is also mapped into a 0.25° spatial resolution in the same spatial extent of the remaining products.

3. Methodology

The flowchart of the developed blending approach consists of three steps and is illustrated in Fig. 2. The three steps are implemented to correct biases of primary data and blended precipitation occurrence and intensity.

[Insert Fig. 2 about here]

3.1 Stage 1: Bias correction of satellite/reanalysis products

The precipitation occurrence and intensity of four primary products are individually adjusted with the gauging observations by a hybrid approach with incorporation of the local intensity scaling (LOCI) and Ratio Bias Correction (RBC) techniques. The LOCI method (Schmidli et al., 2006), which has been widely employed in correcting climate model outputs (e.g., Yin et al., 2020), is introduced here to initially correct the precipitation occurrence of satellite/reanalysis estimates. To implement the LOCI method, a wet threshold is defined as 1 mm/day following Dinku et al. (2008) and Jiang et al. (2019), and then the wet-day occurrence for each gauge is estimated. It is informative to note that for those grids which contain two or more gauges, the observational precipitation series are represented by averaging those gauge records. For the sake of having the same rainy events as observation data, a new wet-day threshold is estimated from the satellite/reanalysis products for each gauge. If the new wet-day threshold is larger than 1mm/day, the intensity of those events lower than the estimated wet-day threshold was rescaled to below 1 mm/day. If the new wet-day threshold is smaller than 1 mm/day, those events with intensity ranging between the new wet-day threshold and 1 mm/day were all rescaled to 1mm/day. More details about the LOCI method can be found in Schmidli et al. (2006) and Yin et al. (2020). The LOCI method is implemented in different months for each gauge, and therefore each satellite/reanalysis dataset is guaranteed to have the same monthly precipitation occurrence as observations at all gauges.

The RBC method (Bhatti et al., 2016) is then employed to cope with the systematic biases of precipitation intensity in a monthly moving window. The correction factor $\beta_{j,m}^{s}$ for each gauge j and month m for the s^{th} ($s=1,2,3,4$) satellite/reanalysis precipitation dataset is calculated by dividing the accumulated gauging observations with estimates in the corresponding grid cell:

$$\beta_{j,m}^{s}=\sum_{i=1}^{N}P_{j,d,m}^{g}\Big/\sum_{i=1}^{N}P_{j,d,m}^{s} \tag{1}$$

where the subscript d and N refer to a specific day and total number of measurements in the m month, respectively; the $P_{j,d,m}^{g}$ and $P_{j,d,m}^{s}$ are gauging observation and estimate of the s^{th} satellite/reanalysis products after LOCI adjustment, respectively.

The bias-corrected precipitation outputs for each product are calculated by multiplication of the primary estimates by the correction factor $\beta_{j,m}^{s}$ as follows:

$$P_{j,d}^{s*}=\beta_{j,m}^{s}\times P_{j,d}^{s} \tag{2}$$

where $P_{j,d}^{s*}$ denotes the adjusted daily precipitation series by the RBC method.

3.2 Stage 2: Merging precipitation occurrence

The primary goal of the 2nd stage is to blend the four bias-corrected products to eliminate missing or false detection of precipitation, which is achieved by optimizing weights that enable to measure the wet/dry event detection capacity of individual members. The “state weights” used for blending wet/dry condition for each day are defined in this stage, which are estimated by measuring the state detection capacity of different products by comparing with observational gauging references. For each gauge j, a score function $F_{j,d}^{s}$ at the d^{th} day for the s^{th} bias-corrected satellite/reanalysis product is defined as follows:

$$F_{j,d}^{s} = (P_{j,d}^{s*} - R_{o}) \cdot (P_{j,d}^{g} - R_{o}) + L(P_{j,d}^{s*} - R_{o}) \cdot L(P_{j,d}^{g} - R_{o}) \tag{3}$$

where R_{o} denotes the precipitation threshold, 1 mm/day in this study; $L(u)$ is an indicator function, if $u \geq 0$, $L(u) = 1$; otherwise $L(u) = 0$.

Then, the state weights could be estimated as follows:

$$\omega_{1,s}(d) = \begin{cases} H(F_{j,d}^{s}) \Big/ \sum_{s=1}^{4} H(F_{j,d}^{s}) & ,\text{if } \sum_{s=1}^{4} H(F_{j,d}^{s}) > 0 \\ 0.25 & ,\text{if } \sum_{s=1}^{4} H(F_{j,d}^{s}) = 0 \end{cases} \tag{4}$$

where $\omega_{1,s}(d)$ denotes the weights of the s^{th} bias-corrected satellite/reanalysis precipitation product at day *d*; $H(F_{j,d}^{s})$ is an indicator function, if $F_{j,d}^{s} > 0$, $H(F_{j,d}^{s}) = 1$;otherwise $H(F_{j,d}^{s}) = 0$; the sum of all the weights $\omega_{1,s}(d)$ in Eq. (4) is equal to 1.

Finally, the precipitation $I_{j,d}$ at day *d* in this stage could be estimated as:

$$I_{j,d} = \sum_{s=1}^{4} \omega_{1,s}(d) \cdot P_{j,d}^{s*} \tag{5}$$

After obtaining the $I_{j,d}$ in blending schemes, we can identify the day *d* as a wet or dry condition by comparing the estimate with the wet threshold. This blending method is implemented for each gauge in a daily moving window in consideration of spatiotemporal variations.

3.3 Stage 3: Merging precipitation intensity

After determining a wet/dry state in the 2nd stage, we define the “intensity weights” in the 3rd stage for estimating precipitation intensity. In this stage, a heuristic blending algorithm based on the CS is developed, and the dynamic BMA method is also employed for the purpose of a comparison. The choice of appropriate training data for calculating dynamic weights is highly dependent upon specifics of algorithms and regions (Sloughter et al., 2007). Unlike previous studies (Ma et al., 2018 a,b; Rahman et al., 2020) selecting the previous 40 days (and the same time from the previous two years) as training period, we find that using data during the previous 55 days and subsequent 55 days for calibration leads to a better estimation accuracy (Fig. S1). Therefore, 110 days in total, are selected to optimize the "intensity weights" for each day, and this optimization procedure is implemented by shifting the training windows day by day at each gauge. In the 3rd stage, the intensity blending is only performed when the estimated precipitation $I_{j,d}$ at the 2nd stage outweighs 1mm/day, and we only select data in wet days (determined by gauging information) in the algorithm for calibration.

3.3.1 Heuristic blending algorithm based on cuckoo search

To optimize the weights at gauge *j* for blending precipitation intensity, the sum of square error (SSE) is selected as an objective function and minimized at each day as follows:

$$\begin{aligned} &min \quad SSE_{j,d} = \sum_{day=d-N}^{d+N} \left\{ \left[\sum_{s=1}^{4} w_s(d) \cdot P_{j,day}^{s*} \right] - P_{j,day}^{g} \right\}^2 \\ &st. \quad \sum_{s=1}^{4} w_s(d) = 1 \\ &0 \le w_s(d) \le 1, \quad s = 1,2,3,4 \end{aligned} \tag{6}$$

where $w_s(d)$ denotes the "intensity weights" of the s^{th} product at day *d*, and the merging precipitation intensity at day *d* is estimated by $I_{j,d}^{*} = \sum_{s=1}^{4} w_s(d) \cdot P_{j,d}^{s*}$; *N* denotes the data size before (or after) day *d* used for calibration (in the case of 55).

The CS algorithm, a powerful and versatile tool for solving nonlinear global optimization problems, is adopted to optimize the $\omega_s(s)$ in Eq. (6). The CS algorithm developed by Yang and Deb (2009) is a bio-inspired metaheuristic algorithm, which has been proved to be more efficient than particle swarm optimization and genetic algorithms in parameter optimizations. The CS was inspired by the obligate brood parasitism of some cuckoo species by laying their eggs in the nests of host birds (Valian et al., 2013). Some cuckoos have evolved in such a way that female parasitic cuckoos can imitate the colors and patterns of the eggs of a few chosen host species, which reduce the probability of the eggs being abandoned and thus increase their re-productivity (Yang and Deb, 2018). For simplicity in implementing the CS, we follow three idealized rules: (1) each cuckoo lays one egg at a time and dumps it in a random nest; (2) the best nests with high quality eggs will be carried over to the next generations; (3) the number of available host nests is fixed, and each host has a randomly generating probability $P_a (P_a \in (0,1))$ to discover an alien egg. In this case, the host bird can either get rid of the egg, or simply abandon the nest and build a completely new nest at a new location.

The CS algorithm uses a balanced combination of a local random walk and the global explorative random walk, controlled by the switching parameter P_a. The local random walk can be expressed as:

$$x_k^{t+1} = x_k^t + \beta z \otimes H(p_a - \varepsilon) \otimes (x_l^t - x_m^t) \tag{7}$$

where x_l^t and x_m^t randomly disturb a sequence of numbers; $H(\cdot)$ denotes a Heaviside function; ε is a random number drawn from a uniform distribution; z and β represent the step size and scaling factor, respectively; $\otimes$ is an entry-wise operation.

Host k of a global random walk is carried out by using Lévy flights as follows:

$$x_k^{t+1} = x_k^t + \alpha \otimes L\grave{e}vy(z,\lambda) \tag{8}$$

$$L\grave{e}vy(z,\lambda) \sim \frac{\lambda\Gamma(\lambda)\sin\left(\frac{\pi\lambda}{2}\right)}{\pi} \cdot \frac{1}{z^{1+\lambda}} \tag{9}$$

where the '~' indicates that the random numbers $L\grave{e}vy(z,\lambda)$ should be drawn from the Lévy distribution that is approximated by a fat-tailed distribution such as a power-law distribution with an exponent λ. The parameter $\alpha > 0$ is the step size scaling factor determined by the scales or bound ranges.

3.3.2 Dynamic Bayesian model averaging scheme

The dynamic BMA scheme was also performed to blend multi-source precipitation estimates by adjusting the posterior probability density function (PDF) to obtain a good fit to the gauging measurements. The PDF of posterior distribution in BMA is represented as:

$$p(I \mid G) = \sum_{s=1}^{S} p(f_s \mid G) \cdot p_s(I \mid f_s, G) \tag{10}$$

where I and G denote the blended precipitation intensity and observation, respectively; S is the number of satellite/reanalysis products; $p(f_s \mid G)$ is the posterior probability of bias-corrected satellite/reanalysis precipitation estimates, also known as the likelihood of ensemble members, with f_s denoting the precipitation estimates of s^{th} member; $p_s(I \mid f_s, G)$ is the posterior distribution of I given estimated f_s and observed G.

The $p(f_s \mid G)$ measures the capacity of bias-corrected satellite/reanalysis products in capturing observed data. After substituting by the weights w_s, Eq. (10) could be expressed as:

$$p(I \mid G) = \sum_{s=1}^{S} w_s \cdot p_s(I \mid f_s, G) \tag{11}$$

The posterior mean $E(I \mid G)$ and variance $Var(I \mid G)$ of the BMA estimation could be expressed as follows:

$$E(I \mid G) = \sum_{s=1}^{S} w_s \cdot f_s \tag{12}$$

$$Var(I \mid G) = \sum_{s=1}^{S} w_s \cdot \left[f_s - E(I \mid G) \right]^2 + \sum_{s=1}^{S} w_s \sigma_s^2 \tag{13}$$

where σ_s^2 is the variance associated with satellite/reanalysis estimates f_s with respect to observation *G*.

The Box-Cox transformation is employed before calibrating the BMA model to follow the Gaussian assumption of the conditional probability distribution $p_s(I \mid f_s, G)$. To achieve a more efficient computation, the log likelihood function is preferred and expressed as follows:

$$LL(w_s, \sigma_s, s = 1, 2, \cdots, S) = \log\left[\sum_{s=1}^{S} w_s \cdot p_s(I \mid f_s, G) \right] = \log\left[\sum_{s=1}^{S} w_s \cdot g(I \mid f_s \cdot \sigma_s^2) \right] \tag{14}$$

where $g(\cdot)$ denotes Gaussian distribution. The Expectation-Maximization algorithm (McLachlan and Krishnan, 2007) is employed to optimize the BMA parameters by maximizing the log likelihood function.

3.4 Precipitation mapping and cross-validation experiments

The important role of blending multi-source precipitation products is to provide useful information in data-sparse or ungauged regions, which is also the primary motivation of this study. The correction factors and optimal dynamic weights in three stages of all gauges were interpolated to the entire mainland China using the OK method at a daily scale. The interpolated weights of the ensemble members were normalized to ensure that their sum is equal to 1. Finally, the blended precipitation estimates were calculated for each grid cell based on the individual data and corresponding correction factors in LOCI and RBC methods as well as optimal grid weights over mainland China.

The performance of the developed blending algorithms for mapping precipitation in ungauged areas is systematically evaluated using a cross validation experiment, i.e., K-fold, implemented separately in eight sub-regions (Fig. 1). In the K-fold (K=10 in this study) experiment, all the gauges were randomly split into ten partitions, and 90% of sites are selected as training and the remaining 10% gauges are used for validation. To fully consider the variations of gauge distributions, the K-fold experiment was repeated 100 times. For each experiment, the parameters of training gauges were extracted and interpolated to the validation sites by the OK method. After normalizing the transferred weights, the blended precipitation over those validation gauges could be estimated. In this way, the validation sites could be considered as independent sites to evaluate the precipitation estimation accuracy of the blending algorithms in ungauged regions.

3.5 Hydrological modelling and performance evaluation

After mapping precipitation over mainland China by blending algorithms, different gridded precipitation products were forced in hydrological models to evaluate their performance in streamflow simulations. Given that the 238 studied catchments differ in climate patterns and underlying surface conditions, we employed four different lumped conceptual hydrological models as candidates: the HMETS model (Martel et al. 2017), GR4J model (Perrin et al., 2003), HBV model (Bergström and Forsman, 1973), and Xinanjiang model (Zhao et al., 1992). These four models have varying numbers of parameters, model structures and runoff yielding mechanisms in modelling rainfall-runoff processes. After generating daily basin average meteorological series from gridded precipitation (primary and blended products) and observational temperature data using the Thiessen polygons method, they were forced to drive the hydrological models for each catchment. A cross-validation approach (Arsenault et al., 2017) was employed for model evaluation, in which the model is calibrated on odd years, whereas it is validated based on even years. As many studied catchments over China are located at data-scarce regions, this study followed numerous previous studies (e.g., Tobin and Bennett, 2010; Ma et al., 2018b), and recalibrated models by using different primary and blended precipitation datasets. All setups use the same model forcing except precipitation, and thus differences between the model efficiency can represent the differences in precipitation datasets (Jiang et al., 2019).

We optimized the parameters of the four hydrological models by using the Shuffled Complex Evolution (SCE-UA) method developed at the University of Arizona (Duan et al., 1992). The SCE-UA algorithm integrates the advantages of several effective global optimization medthods. This method employs both deterministic search strategies and random schemes to achieve a relatively efficient search capacity. The Kling-Gupta efficiency (*KGE*) is selected as the objective function and is maximized during calibration (Gupta et al., 2009):

$$KGE = 1 - \sqrt{(r-1)^2 + (\zeta - 1)^2 + (\vartheta - 1)^2} \quad (15)$$

where r and ξ (or ϑ) refer to the Pearson's correlation coefficient and ratio of standard deviation (or mean index) of observation and simulations, respectively.

4. Results

4.1 Bias correction performance

To assess the bias correction performance in the 1st stage, eight statistical metrics are used to evaluate estimation accuracy of precipitation intensity (RMSE, MAE, CC and KGE) and occurrence (POD, FAR, CSI and HSS), respectively. These metrics are expressed in Table 1, and the last four indices are calculated from a 2×2 contingency matrix composed of four parameters, of which *a* is the number of rainfall events successfully detected; *b* is the number of dry events erroneously identified as rain events (false alarms); *c* is the number of missing events; and *d* is the sum of events that are neither detected nor observed. HSS measures the accuracy of the estimates accounting for matches due to random chance, and is different from the POD, FAR and CSI, which are highly affected by the climatology of different study regions.

[Insert Table 1 about here]

The primary precipitation products show substantial biases over mainland China (Table 2). For instance, the daily statistical metrics range from 7.17-8.69 mm, 0.25-0.49 and 0.24-0.44, in terms of RMSE, CC and HSS, respectively. The TMPA 3B42V7 generally performs worst with the highest RMSE at 8.69 mm and the lowest CC of 0.25, while the ERA5 exhibits the preferable performance in terms of the lowest RMSE (and MAE, FAR) at 7.17 mm (and 2.78 mm, 0.52) and the highest CC at 0.49. It is difficult to determine a best precipitation product in terms of all evaluation metrics. For example, the IMERG Final exhibits highest KGE among the four products while it performs worse than the ERA5 in terms of other metrics. After bias correction by combining the LOCI and RBC methods, the adjusted products show great improvements. For example, the RMSE and MAE of the bias-corrected products correspondingly decrease to 6.61-7.92 mm (by 2.8-7.8%) and 2.40-3.11 mm (by 1.3-13.7%), respectively. Moreover, the CC index of the four products improves to 0.28-0.55.

[Insert Table 2 about here]

To systematically validate the seasonal performance of the bias correction method, the seasonal average statistical results of RMSE and CC (MSE and HSS) metrics for primary and adjusted precipitation products are presented in Fig. 3 (Fig. S2). After bias correction, the seasonal statistical metrics exhibit substantial improvement in all seasons over mainland China. We also examined the potential heterogeneity of bias-correction performance for eight sub-regions over China (Fig. 4). South China generally shows a poorer performance in terms of the evaluation metrics, whereas most products achieve well estimation scores in Northwest China. More importantly, the statistical metrics in all sub-regions have been generally improved after bias correction for the four precipitation products. All the above results demonstrate a good bias correction performance in the 1st phase, verifying its suitability and reliability in reducing estimation errors of four involved products over mainland China.

[Insert Fig. 3 about here]

[Insert Fig. 4 about here]

4.2 Spatiotemporal distribution of dynamic weights

The four bias-corrected precipitation products are employed to train the flexible three-stage dynamic blending schemes, and both the CS algorithm and BMA method are considered in the last phase. Fig. 5 shows the spatial distribution of average annual weights for the 2nd and 3rd stages over mainland China, revealing that the weights of different members significantly vary from stages and algorithms. Particularly, the weights used for determining a wet/dry day and for estimating precipitation intensity show substantial differences. The weights of four bias-corrected products in 2nd stage generally range from 20-35% across China, while in the 3rd stage the weights of some members can dominate over 50% and the TMPA 3B42V7 might account for below 10% across the majority of landmasses (Fig. 5). This significant discrepancy also highlights the necessity of determining wet/dry state and precipitation intensity in different stages. Beside the differences in two stages, the weights of three-stage CS algorithm and dynamic BMA method also show substantial variations over most regions. For instance, the ERA5 dataset accounts for over 45% in most areas under the CS-based blending scheme, while the weights are generally 20-35% for the dynamic BMA method. Moreover, the weights are accompanied by certain spatial heterogeneity. For instance, the weights of PERSIANN-CDR in Northwest China vary from 10-15%, which are far lower than those in other regions (generally accounting for 20-25%).

[Insert Fig. 5 about here]

To further investigate the temporal distribution of weights, the spatial average weights in eight sub-regions are plotted against the day of year (DOY). Comparing with the 3rd blending stage (Figs. 6 and S3), the 2nd stage shows less seasonal variability in dynamic weights (Fig. S4). In the three-stage CS blending scheme, the eight sub-regions are typically dominated by EAR5, and they are accompanied by considerable spatial and temporal variabilities. For example, the relative weights in different seasons did not change much in North China, while a strong seasonal variability is detected in South China, i.e., the ERA5 accounts for almost 60 percent in winter and the weights decline to 40 percent in summer (Fig. 6). The TMPA 3B42V7 consistently shows weaker skills and thus receives lower weighting scores (around 10%) throughout the year, which holds true in all sub-regions. It implies that the TMPA algorithm needs to be updated with an effective ground observation network. Surveying the existence of the relationship between satellite/reanalysis precipitation with regard to higher weights and the specific type of climate pattern would be beneficial. However, the temporal distribution of weights might be not directly related to climate dynamics across mainland China, as the weights are more highly governed by different precipitation members.

[Insert Fig. 6 about here]

4.3 Performance assessments of different blending schemes

The developed three-stage CS algorithm and dynamic BMA schemes are employed to reorganize the precipitation regimes by merging the bias-corrected members at all gauges over mainland China, respectively. To test the benefits of the three stages, we also compared their performance with two blending schemes without the 2nd stage, i.e., omitting the wet/dry event detection procedure. Table 2 summarizes the average daily metrics of the individuals and four ensemble methods over mainland China. The four blended data generally perform better than both primary and bias-corrected members in terms of evaluation metrics. For instance, the MAE of the four blended ensembles ranges from 2.05-2.63 mm, decreased by 26.3-36.7% (14.5-31.9%) compared with the original (adjusted) datasets (Table 2). Comparing the two-stage and three-stage schemes, it is informative to notice that the three-stage blending approaches achieve better scores than the two-stage methods in terms of all eight metrics. For example, the CC and POD value of the two-stage methods (0.47-0.53 and 0.74-0.76) is improved to 0.51-0.61 and 0.86 in the three-stage schemes, respectively. The CSI and HSS increase by 92% and 100%, and the FAR shows a substantial reduction to 0.15 comparing with the 0.55 of the two-stage approaches. These results further highlight the necessity of carefully incorporating the wet/dry event detection phase into the precipitation blending schemes.

The spatial distributions of eight statistical error metrics for different products are presented in Figs. 7-8 and Figs. S5-S10. In the three-stage framework, as the precipitation occurrences are determined in the second stages, the CS algorithm and dynamic BMA method perform equally in detecting event states in terms of POD, FAR, CSI and HSS. The three-stage CS algorithm generally achieve superior scores than the dynamic BMA method in terms of RMSE, MAE, CC and KGE. To further investigate impacts of time scales on precipitation estimate accuracy, focusing on four sub-regions as examples, density scatters of estimated and observed areal precipitation are plotted at both ten-day and monthly scales (Figs. 9 and S11-13). Clearly, all products show stronger correlation with rain gauges at longer time scales. Moreover, the blended precipitation under three-stage CS algorithm consistently shows the highest CC and lowest RMSE metrics, and its estimates agree well with the gauge-based observations with a correlation coefficient over 0.98. Overall, the three-stage CS algorithm demonstrates the best scores in terms of all eight metrics at the calibrated gauges, highlighting the feasibility of using heuristic three-stage schemes to merge multi-source precipitation data over mainland China.

[Insert Fig. 7 about here]

[Insert Fig. 8 about here]

[Insert Fig. 9 about here]

To further evaluate the potential benefits of the blended precipitation products in ungauged regions, the K-fold experiment is used to randomly select gauges for model calibration and validation. After randomly splitting all the gauges into 10 groups for each sub-region, the sites of nine groups are organized to train the blending schemes while the remaining sites are treated as validation. The parameters of trained sites are interpolated by the OK method to the validated gauges. After normalization of transferred weights, the blended precipitation series under different schemes are estimated for those sites. The random sampling is repeated 100 times to guarantee robustness of the results, and the average results of daily metrics of the primary and four blended precipitation products over random validated gauges are presented in Table 3. As for the blended results over verified gauges in Table 3, the precipitation estimate skills are slightly worse than those obtained in Table 2. This is reasonable as the blended precipitation at randomly chosen validation gauges are estimated by transferred parameters from surrounding calibrated sites, rather than fitting the ensemble model with their located data.

[Insert Table 3 about here]

In Table 3, the four blended estimates all show better performance compared to the raw multi-satellite/reanalysis products. Statistically, the averaged values of RMSE, MAE and CC for the four primary datasets range from 6.51-9.59 mm, 2.49-3.36 and 0.25-0.48, respectively, while the blended products improve the estimating scores by 10-43% (Table 3). The four blended products also show better scores than primary precipitations in terms of the other evaluation indexes. Moreover, the three-stage CS scheme shows best performance among the four blending schemes. For instance, the CC and HSS under such three-stage heuristic algorithm are 0.50 and 0.54, respectively, while the two-stage schemes yield these metrics as 0.44-0.47 and 0.40-0.42, respectively. Fig. 10 presents the evaluation results over eight sub-regions in terms of all metrics, which consistently prove the best performance of the three-stage CS blending algorithm in comparison with four primary datasets. These comparisons highlight the superiority of using the blending method to merge multi-satellite/reanalysis precipitation data and prove that the three-stage CS scheme outperforms both two-stage methods and dynamic BMA for multiple data fusion over mainland China.

[Insert Fig. 10 about here]

4.4 Precipitation mapping and hydrological performance assessment

Based on the bias correction parameters and dynamic weights under the three-stage CS algorithm, a new blended dataset covering 2004-2018 is produced by interpolation and normalization with information of all gauges at a daily 0.25° scale over mainland China. To consider the different climatic pattern and underlying surface conditions for different catchments, the HMETS, GR4J, HBV and Xinanjiang models are forced by the blended data over the 238 studied catchments, and the model with the largest KGE value is selected for hydrological simulations in each basin (Fig. S14). The Xinanjiang model performs best over a majority of China's catchments, while the GR4J model exhibits best simulation performance in 25 percent of basins.

[Insert Fig. 11 about here]

The best performing hydrological model at each catchment is also forced by four primary multi-satellite/reanalysis members, and the KGE values of different forcing schemes during calibration and validation periods are presented in Fig. 11 and Fig. S15, respectively. Among the four primary precipitation datasets, the hydrological models achieve a relatively better performance when forcing by the EAR5 and IMERG-Final precipitation estimates during calibration period, with the KGE ranging from 0.4-0.6 in most catchments. The 3B42V7 and PERSIANN-CDR exhibit relatively worse hydrological skill in terms of a lower KGE value, with only 27% and 30% catchments exhibiting a satisfactory KGE (>0.5). Comparing the primary and blended datasets, the mapped blended precipitation estimates could significantly improve the hydrological performance, which is supported by a higher KGE values over the majority of catchments. Particularly over humid and semi-humid regions, the blended product could generally improve the KGE from 0.3-0.6 to about 0.6-0.9, with an increasing rate of 12-35%. When forced by the blended data, almost all studied catchments exhibit a KGE value larger than 0.6 during calibration period, and 62 percent of catchments yielded a KGE higher than 0.8 (Fig. 11). Although we observe relatively low KGE values in a few basins during the validation period, most measured basins still had satisfactory KGE (>0.6) and 43% of catchments yielded a KGE higher than 0.75 (Fig. S15). Overall, the new blended precipitation datasets could substantially facilitate hydrological modelling, implying its important role to serve as an alternative in representing hydro-climatic transferability over mainland China, particularly in those data-sparse regions.

4.5 Performance comparison with MSWEP V2 dataset

In order to comprehensively understand the strengths and weaknesses of the new blended precipitation dataset, a state-of-the-art high-quality merged precipitation product, Multi-Source Weighted Ensemble Precipitation Version 2 (MSWEP V2), was used for comparison. MSWEP is developed to provide globally 3-hour precipitation data at 0.25° spatial resolution from 1979 to 2017, and the latest Version 2 was released by Beck et al. (2019). The spatial distribution of eight evaluation metrics for MSWEP V2 product is presented in Fig. 12, and the average daily metrics over mainland China is also summarized in Table 2. Results show that MSWEP V2 generally achieves better skills than primary four satellite/reanalysis datasets. MSWEP V2 also exhibits higher estimation scores than the four bias-corrected members except for the adjusted ERA5 dataset. However, when comparing with our new blended dataset, MSWEP V2 generally perform worse in terms of all evaluation metrics. For instance, the HSS of MSWEP V2 (0.47) is much lower than that of the blended dataset (0.82). We also compared the performance of MSWEP V2 and blended datasets at different sub-regions and seasons (see Figs. 3-4 and Fig. S2), further confirming the superior performance of our blended dataset to MSWEP V2. In order to compare the performance of streamflow simulation, MSWEP V2 dataset is used to force the best-performing hydrological model for each basin. The KGE values of streamflow simulation during calibration and validation periods are also demonstrated in Fig. 11 and Fig. S15, respectively. When forced by MSWEP V2 dataset, about 70 percent of catchments have KGE values lower than 0.55, and only very few catchments achieve KGE as high as 0.7 in calibration period. MSWEP V2 may show better hydrological performance than ERA5 in limited catchments, but it is far worse than the blended product over a majority of China’s catchments. Ma et al. (2018b) also supported our finding that MSWEP V2 usually shows substantial biases in precipitation estimation and hydrological utilization in a plateau region of China. This can be partly attributed to the differences of the employed bias correction and blending methods. More importantly, MSWEP V2 focused on the globally gridded precipitation reconstruction, and only a small portion of in-situ gauges were used for MSWEP V2 data over China due to the data constraint (Beck et al., 2019). This study fully takes advantage of the gauge data from China Meteorological Administration, which is important for improving precipitation estimation quality of multi-source data blending.

5. Discussion

In spite of superior performance of the merging algorithms, some work still needs to be further investigated. Future investigations might be devoted to produce higher-resolution datasets by incorporating the underlying physical mechanisms of precipitation generation in the multi-source data fusion frameworks. For instance, precipitation intensity is highly determined by atmospheric temperatures and relative humidity as governed by the Clausius-Clapeyron relationship (Yin et al., 2018). As the atmospheric reanalysis dataset (e.g., the ERA5 used in this study) can provide hourly climate variables representing both energy and water flux states, the daily precipitation intensity may be temporally distributed into a sub-daily scale by utilizing information from hourly temperature and humidity variables. To provide precipitation reference in ungauged region over China, this study aims to produce a high-quality retrospective dataset, thus employing a best-training scheme for calibrating the dynamic blending approach (Fig. S1). In such training scheme, the daily weights are calculated using data before and after the specific day, which only works for blending retrospective datasets. To test potential usefulness of the dynamic blending approach, taking the 2018-year as an example period, we also used the previous 40 days (and the same time from the previous two years) as training period following Rahman et al. (2020). The average daily metrics of the primary and the blended precipitation dataset in 2018-year over mainland China is presented in Table S1. Results show that the dynamic blending method can still substantially reduce the biases of primary precipitation dataset when trained by only using past observations. This study did not take near-real time precipitation datasets as blending candidates, which may limit potential application in monitoring and forecasting. However, comprehensive assessments suggest that the three-stage blending method provides a useful precipitation dataset for data-sparse regions, which is important for water resources management and planning over China.

The developed three-stage blending algorithm is a statistical-based method, which is accompanied by estimation uncertainty sourced from primary data, model structure and parameter estimations. Although we attempted to correct systematic bias of four individual products, only one hybrid approach incorporating LOCI and RBC methods is employed. Considering the further transferability of correction factors by OK interpolation, we did not examine more sophisticated bias-correction algorithms. However, numerous approaches may work such as daily translation (Yin et al., 2020), cumulative distribution function matching (Mastrantonas et al., 2019), copula-based correction (Sharifi et al., 2019) and stepwise regression method (Lu et al., 2019). Future work should be focused on further comparing and evaluating the adjusting performance of different bias correction methods in reducing errors of primary datasets. While this study employed a dynamic blending method to produce a deterministic dataset (i.e. a single "best guess" realization of precipitation), there is parallel research that focuses on developing probabilistic precipitation estimates. A probabilistic design of satellite/reanalysis precipitation products is also important for hydrological application, because it provides the possible range of estimates. For example, Kirstetter et al. (2018) proposed a new method, PIRSO (Probabilistic Precipitation Estimation using InfraRed Satellite Observations), to estimate probabilistic precipitation rates with space-based infrared sensors. Although it is challenging for this study to release a probabilistic dataset at a national scale, we still attempt to characterize the hydrological simulation uncertainties when forced by satellite/reanalysis precipitation. Following the general concept of a probabilistic estimation framework, four primary precipitation datasets were individually run through the Xinanjiang model. The precipitation is treated as a 4-member ensemble process rather than a deterministic one, and then the four simulated streamflow members is blended by the dynamic BMA method. The streamflow simulation performances of different schemes are presented in Fig. 13 and Fig. S16. When treating the precipitation as a 4-member ensemble, most catchments show higher KGE values than those schemes when forced by individual primary datasets. This finding supports that the ensemble-based approach may provide more reliable information for hydrological simulation. Among all the considered six calibration schemes, the model forced by blended precipitation dataset usually shows best performance in almost all catchments. As a result, it is better to reduce the precipitation biases before driving hydrological models, which also suggests the potential usefulness of our blended precipitation dataset in streamflow simulation. Here, it is difficult to detect and systematically eliminate different uncertainty components; therefore, removing the integrated uncertainty and further improving precipitation blending accuracy have to be further researched.

6. Conclusions

This study develops a three-stage framework to integrate three multi-satellite precipitation datasets (IMERG Final, TMPA 3B42V7 and PERSIANN-CDR), a latest atmospheric reanalysis product ERA5 and gauge dataset, particularly to provide blended precipitation estimates in data-sparse or ungauged regions. This framework can simultaneously correct precipitation occurrence and intensity, and is performed to produce a new precipitation dataset at a daily 0.25° grid scale over mainland China. The developed method is systematically evaluated in terms of eight statistical metrics in both gauged sites, and is also evaluated at randomly sampled sites by K-fold experiments. The hydrological performance of blended and primary multi-satellite/reanalysis members are also evaluated by forcing four hydrological models in 238 catchments. The main conclusions are summarized as follows.

(1) The four primary precipitation products show substantial biases over mainland China, and ERA5 exhibits the best performance in terms of most error evaluation metrics. After bias correction by combining the LOCI and RBC methods, the adjusted products show significant improvements in both capturing precipitation occurrence and intensity, generally with an improving rate of 2.8-13.7% after adjustment.

(2) The three-stage blending approaches achieve better scores than the two-stage methods and individual members, and the CS algorithm generally performs superior

estimation skills than the dynamic BMA method. The K-fold experiments also proved that blended products can improve the estimating scores by 10-43%, implying substantial benefits of precipitation blending in ungauged regions.

(3) The mapped blended precipitation estimates could significantly improve the hydrological performance in comparison with primary members, with improvement of the KGE values of simulated streamflow by 12-35% in most catchments over mainland China. Overall, the developed three-stage heuristic method enables facilitating hydrological modelling, and therefore may play an important role in hydro-climatological applications over data-sparse regions in mainland China.

Acknowledgements

This work was funded by the National Natural Science Foundation of China (52009091; 51879192) and the Natural Science Foundation of Hubei Province (NO. 2020CFB239; 2020CFB132). This work was supported by the China Postdoctoral Science Foundation (2020M682478) and Post-doctoral Innovative Talent Support Program of China (BX20200257). The study is also funded by "111 Project" Fund of China (B18037), and is partly funded by the Ministry of Foreign Affairs of Denmark and administered by Danida Fellowship Centre (File number: 18-M01-DTU).

References

Amjad, M., Yilmaz, M. T., Yucel, I., Yilmaz, K. K., 2020. Performance evaluation of satellite- and model-based precipitation products over varying climate and complex topography. J. Hydrol. 584, 124707.

Arsenault, R., Essou, G. R. C., Brissette, F. P., 2017. Improving hydrological model simulations with combined multi-input and multi-model averaging frameworks. J. of Hydrol. Eng. 22, 04016066.

Ashouri, H., Hsu, K. L., Sorooshian, S., Braithwaite, D. K., Knapp, K. R., Cecil, L. D., Nelson, B. R., Prat, O. P., 2015. PERSIANN-CDR: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies. B. Am. Meteorol. Soc. 96(1), 69-83.

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., van Dijk, A. I., McVicar, T. R., Adler, R. F., 2019. MSWEP V2 global 3-hourly 0.1° precipitation: methodology and quantitative assessment. B. Am. Meteorol. Soc. 100, 473-500.

Bergstrom, S., & Forsman, A., 1973. Development of a conceptual deterministic rainfall-runoff model. Hydrol. Res. 4(3), 147-170.

Beria, H., Nanda, T., Singh Bisht, D., Chatterjee, C., 2017. Does the GPM mission improve the systematic error component in satellite rainfall estimates over TRMM? An evaluation at a pan-India scale. Hydrol. Earth Syst. Sci. 21, 6117-6134.

Bharti, V., Singh, C., 2015. Evaluation of error in TRMM 3B42V7 precipitation estimates over the Himalayan region. J. Geophys. Res.-Atmos. 120, 12458-12473.

Bhatti, H. A., Rientjes, T., Haile, A. T., Habib, E., Verhoef, W., 2016. Evaluation of bias correction method for satellite-based rainfall data. Sensors 16, 884.

Bhuiyan, M. A., Nikolopoulos, E. I., Anagnostou, E. N., 2019. Machine learning-based blending of satellite and reanalysis precipitation datasets: A multiregional tropical complex terrain evaluation. J. Hydrometeorol. 20, 2147-2161.

Chao, L., Zhang, K., Li, Z., Zhu, Y., Wang, J., Yu, Z., 2018. Geographically weighted regression based methods for merging satellite and gauge precipitation. J. Hydrol. 558, 275-289.

Chen, G., Sha, W., Iwasaki, T., 2009. Diurnal variation of precipitation over southeastern China: 2. Impact of the diurnal monsoon variability. J. Geophys. Res.-Atmos. 114, D21105.

Dinku, T., Chidzambwa, S., Ceccato, P., Connor, S. J., Ropelewski, C. F., 2008. Validation of high-resolution satellite rainfall products over complex terrain. Int. J. Remote Sens. 29, 4097-4110.

Duan, Q., Sorooshian, S., Gupta, V., 1992. Effective and efficient global optimization for conceptual rainfall- runoff models. Water Resour. Res. 28(4), 1015-1031.

Graham, R. M., Hudson, S. R., Maturilli, M., 2019. Improved performance of ERA5 in arctic gateway relative to four global atmospheric reanalyses. Geophys. Res. Lett. 46, 6138-6147.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377, 80-91.

Huffman, G. J., Adler, R. F., Arkin, P., Chang, A., Ferraro, R., Gruber, A., Janowiak, J., McNab, A., Rudolf, B., Schneider, U., 1997. The global precipitation climatology project (GPCP) combined precipitation dataset. B. Am. Meteorol. Soc. 78, 5-20.

Huffman, G. J., Adler, R. F., Bolvin, D. T., Nelkin, E. J., 2010. The TRMM Multi-Satellite Precipitation Analysis (TMPA). Satellite Rainfall Applications for Surface Hydrology https://doi.org/10.1007/978-90-481-2915-7_1.

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., 2019. Integrated multi-satellitE retrievals for GPM (IMERG) technical documentation. NASA.

Jiang, L., Bauer Gottwein, P., 2019. How do GPM IMERG precipitation estimates perform as hydrological model forcing? Evaluation for 300 catchments across Mainland China. J. Hydrol. 572, 486-500.

Kirstetter, P.-E., Karbalaee, N., Hsu, K, Hong, Y., 2018. Probabilistic precipitation rate estimates with space-based infrared sensors. Q. J. Roy. Meteor. Soc. 144, 191-205.

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

Li, H., Hong, Y., Xie, P., Gao, J., Niu, Z., Kirstetter, P., Yong, B., 2015. Variational merged of hourly gauge-satellite precipitation in china: preliminary results. J. Geophys. Res.-Atmos. 120, 9897-9915.

Lu, X., Tang, G., Wang, X., Liu, Y., Jia, L., Xie, G., Li, S., Zhang, Y., 2019. Correcting GPM IMERG precipitation data over the Tianshan Mountains in China. J. Hydrol. 575, 1239-1252.

Luo, H., Ge, F., Yang, K., Zhu, S., Peng, T., Cai, W., Liu, X., Tang, W., 2019. Assessment of ECMWF reanalysis data in complex terrain: can the CERA-20C and ERA-Interim datasets replicate the variation in surface air temperatures over Sichuan, China?. Int. J. Climatol. 39(15), 15619-5634.

Ma, Y., Hong, Y., Chen, Y., Yang, Y., Tang, G., Yao, Y., Long, D., Li, C., Han, Z., Liu, R., 2018a. Performance of optimally merged multisatellite precipitation products using the dynamic Bayesian Model Averaging scheme over the Tibetan Plateau. J. Geophys. Res.-Atmos. 123, 814-834.

Ma, Y., Yang, Y., Han, Z., Tang, G., Maguire, L., Chu, Z., Hong, Y., 2018b. Comprehensive evaluation of ensemble multi-satellite precipitation dataset using the dynamic Bayesian model averaging scheme over the Tibetan Plateau. J. Hydrol. 556, 634-644.

Maggioni, V., Meyers, P. C., Robinson, M. D., 2016. A review of merged high resolution satellite precipitation product accuracy during the Tropical Rainfall Measuring Mission (TRMM)-Era. J. Hydrometeorol. 17(4), 1101-1117.

Martel, J. L., Demeester, K., Brissette, F. P., Arsenault, R., Poulin, A., 2017. HMET: a simple and efficient hydrology model forteaching hydrological modelling, flow forecasting and climate change impacts. International Journal of Engineering Education 33(4), 1307-1316.

Massari, C., Maggioni, V., Barbetta, S., Brocca, L., Ciabatta, L., Camici, S., Moramarco, T., Coccia, G., Todini, E., 2019. Complementing near-real time satellite rainfall products with satellite soil moisture-derived rainfall through a

Bayesian inversion approach. J. Hydrol. 573, 341-351.

Massari, C., Brocca, L., Pellarin, T., et al., 2020. A daily 25 km short-latency rainfall product for data-scarce regions based on the integration of the Global Precipitation Measurement mission rainfall and multiple-satellite soil moisture products. Hydrol. Earth Syst. Sci. 24, 2687-2710.

Mastrantonas, N., Bhattacharya, B., Shibuo, Y., Rasmy, M., Espinozadavalos, G., Solomatine, D., 2019. Evaluating the benefits of merging near-real-time satellite precipitation products: a case study in the Kinu basin region, Japan. J. Hydrometeorol. 20, 1213-1233.

McLachlan, G. J., Krishnan, T., 2007. The EM algotithm and extensions. Second Edition. Hoboken, NJ: John Wiley. https://doi.org/10.1002/9780470191613.ch5.

Nogueira, M., 2020. Inter-comparison of ERA-5, ERA-interim and GPCP rainfall over the last 40 years: Process-based analysis of systematic and random differences. J. Hydrol. 583, 124632.

Prakash, S., Mitra, A. K., Aghakouchak, A., Liu, Z., Norouzi, H., Pai, D. S., 2018. A preliminary assessment of GPM-based multi-satellite precipitation estimates over a monsoon dominated region. J. Hydrol. 556, 865-876.

Raftery, A. E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133, 1155-1174.

Rahman, K. U., Shang, S., Shahid, M., Li, J., 2018. Developing an ensemble precipitation algorithm from satellite products and its topographical and seasonal evaluations over Pakistan. Remote Sens. 10, 1835.

Rahman, K. U., Shang, S., Shahid, M., Wen, Y., Khan, Z., 2020. Application of dynamic clustered Bayesian Model Averaging (DCBA) algorithm for merging multi-satellite precipitation products over Pakistan. J. Hydrometeorol. 21, 17-37.

Schmidli, J., Frei, C., Vidale, P. L., 2006. Downscaling from GCM precipitation: a benchmark for dynamical and statistical downscaling methods. Int. J. Climatol. 26, 679-689.

Sharifi, E., Saghafian, B., and Steinacker, R., 2019. Copula-based stochastic uncertainty analysis of satellite precipitation products. J. Hydrol. 570, 739-754.

Shen, Y., Xiong, A., Hong, Y., Yu, J, Pan, Y., Chen, Z., Saharia, M., 2014. Uncertainty analysis of five satellite-based precipitation products and evaluation of three optimally merged multi-algorithm products over the Tibetan Plateau. Int. J. Remote Sens. 35, 6843-6858.

Shi, X., Xu, X., 2007. Regional characteristics of the interdecadal turning of winter/summer climate modes in Chinese mainland. Chinese Sci. Bull. 52, 101-112.

Sloughter, J. M. L., Raftery, A. E., Gneiting, T., Fraley, C., 2007. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. Mon. Weather Rev. 135(9), 3209-3220.

Sunilkumar, K., Narayana Rao, T., and Satheeshkumar, S., 2016. Assessment of small-scale variability of rainfall and multi-satellite precipitation estimates using measurements from a dense rain gauge network in Southeast India. Hydrol. Earth

Syst. Sci. 20, 1719-1735.

Tang, G., Ma, Y., Long, D., Zhong, L., Hong, Y., 2016. Evaluation of GPM Day-1 IMERG and TMPA Version-7 legacy products over Mainland China at multiple spatiotemporal scales. J. Hydrol. 533, 152-167.

Tang, G., Clark, M. P., Papalexiou, S. M., Ma, Z., Hong, Y., 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.

Tong, K., Su, F., Yang, D., et al. 2013. Tibetan Plateau precipitation as depicted by gauge observations, reanalyses and satellite retrievals. Int. J. Climatol. 34(2), 265-285.

Tarek, M., Brissette, F. P., Arsenault, R., 2020. Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America. Hydrol. Earth Syst. Sci. 24, 2527-2544.

Tobin, K. J. and Bennett, M. E., 2010. Adjusting satellite precipitation data to facilitate hydrologic modelling. J. Hydrometeorol. 11, 966-978.

Valian, E., Tavakoli, S., Mohanna, S., Haghi, A., 2013. Improved cuckoo search for reliability optimization problems. Comp. AND Eng. 64, 459-468.

Wright, D.B., Dalia, B.K., Yatheendradas, S., 2017. Satellite precipitation characterization, error modeling, and error correction using censored shifted gamma distributions. J. Hydrometeorol. 18, 2801-2815.

Yang, C., Yuan, H., Su, X., 2020. Bias correction of ensemble precipitation forecasts in the improvement of summer streamflow prediction skill. J. Hydrol. 588, 124955.

Yang, X., Deb, S., 2009. Cuckoo search via Lévy flights, Proceeding of World Congress on Nature and Biologically Inspired Computing (NaBIC). IEEE Publications, Coimbatore, India, USA. 210-214.

Yang, X., Deb, S., 2018. Cuckoo Search: State-of-the-Art and Opportunities. IEEE International Conference on Soft Computing & Machine Intelligence. https://doi.org/10.1109/ ISCMI.2017.8279597.

Yin, J., Gentine, P., Zhou, S., Sullivan, S. C., Wang, R., Zhang, Y., Guo, S., 2018. Large increase in global storm runo extremes driven by climate and anthropogenic changes. Nat. Commun. 9, 4389.

Yin, J., Guo, S., Gu, L., He, S., Ba, H., Tian, J., Li, Q., Chen, J., 2020. Projected changes of bivariate flood quantiles and estimation uncertainty based on multi-model ensembles over China. J. Hydrol. 585, 124760.

Zhang, Q., Xiao, M., Singh, V., Liu, L., Xu, C-Y., 2015. Observational evidence of summer precipitation deficit- temperature coupling in China. J. Geophys. Res.-Atmos. 120, 10040-10049.

Zhao, R., 1992. The Xinanjiang model applied in China. J. Hydrol. 135, 371-381.

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.

List of Tables

Table 1: Summary of the statistical metrics for evaluating the performance of precipitation products.

ID	Metric	Abbreviation	Expression	Perfect score
1	Root mean square error	RMSE(mm)	$\sqrt{\frac{\sum_{j=1}^{N}(S_j - G_j)^2}{N}}$	0
2	Mean absolute error	MAE(mm)	$\frac{\sum_{j=1}^{N}\left\|S_j - G_j\right\|}{N}$	0
3	Correlation coefficient	CC(-)	$\frac{\sum_{j=1}^{N}(S_j - \overline{S})(G_j - \overline{G})}{\sqrt{\sum_{j=1}^{N}(S_j - \overline{S})^2}\sqrt{\sum_{j=1}^{N}(G_j - \overline{G})^2}}$	1
4	Kling-Gupta efficiency	KGE(-)	Eq. (15)	1
5	Probability of detection	POD(-)	$\frac{a}{a+c}$	1
6	False alarm ratio	FAR(-)	$\frac{b}{a+b}$	0
7	Critical success index	CSI(-)	$\frac{a}{a+b+c}$	1
8	Heidke skill score	HSS(-)	$\frac{2(a \cdot d - b \cdot c)}{\left[(a+c)\cdot(c+d)+(a+b)\cdot(b+d)\right]}$	1

Note: G_j (S_j)indicates gauge observation (estimates); *j* and *N* represent time step and total length, respectively; $\overline{G}$ ($\overline{S}$)is the mean value of gauge observations(estimates).

Table 2: The average daily metrics of the primary, adjusted precipitation products, blended ensembles and MSWEP V2 dataset over mainland China.

ID	Products	RMSE(mm)	MAE(mm)	CC	KGE	POD	FAR	CSI	HSS
1	ERA5	7.17	2.78	0.49	0.23	0.69	0.52	0.42	0.44
2	IMERG Final	8.07	2.93	0.39	0.29	0.57	0.53	0.35	0.37
3	TMPA 3B42V7	8.69	3.05	0.25	0.13	0.31	0.54	0.22	0.24
4	PERSIANN-CDR	7.87	3.18	0.33	0.15	0.62	0.63	0.3	0.27
5	ERA5_cor	6.61	2.40	0.55	0.44	0.65	0.44	0.43	0.49
6	IMERG Final_cor	7.8	2.77	0.42	0.41	0.53	0.5	0.36	0.38
7	TMPA 3B42V7_cor	7.92	3.01	0.28	0.25	0.57	0.65	0.28	0.26
8	PERSIANN-CDR_cor	7.65	2.97	0.37	0.32	0.53	0.6	0.31	0.29
9	Two-stage CS	6.53	2.59	0.53	0.35	0.76	0.55	0.39	0.41
10	Two-stage BMA	6.87	2.63	0.47	0.29	0.74	0.55	0.39	0.41
11	Three-stage CS	**6.18**	**2.05**	**0.61**	**0.49**	**0.86**	**0.15**	**0.75**	**0.82**
12	Three-stage BMA	6.68	2.21	0.51	0.34	0.86	0.15	0.75	0.82
13	MSWEP V2	7.17	2.51	0.48	0.34	0.66	0.46	0.42	0.47

Note: The bold text stands for the best score for the comparing members.

Table 3: The average daily metrics of the primary and four blended precipitation products in randomly sampled validation gauges.

ID	Products	RMSE(mm)	MAE(mm)	CC	KGE	POD	FAR	CSI	HSS
1	ERA5	6.51	2.49	0.48	0.23	0.68	0.56	0.39	0.41
2	IMERG Final	7.88	2.50	0.37	0.21	0.55	0.55	0.33	0.36
3	TMPA 3B42V7	9.59	3.36	0.25	0.12	0.48	0.61	0.27	0.28
4	PERSIANN-CDR	6.69	2.70	0.31	0.06	0.60	0.65	0.28	0.26
5	Two-stage CS	6.63	2.62	0.47	0.21	0.66	0.57	0.38	0.42
6	Two-stage BMA	6.92	2.69	0.44	0.21	0.65	0.58	0.38	0.40
7	Three-stage CS	**6.21**	**2.09**	**0.50**	**0.40**	**0.70**	**0.40**	**0.43**	**0.54**
8	Three-stage BMA	6.76	2.23	0.45	0.22	0.70	0.40	0.43	0.54

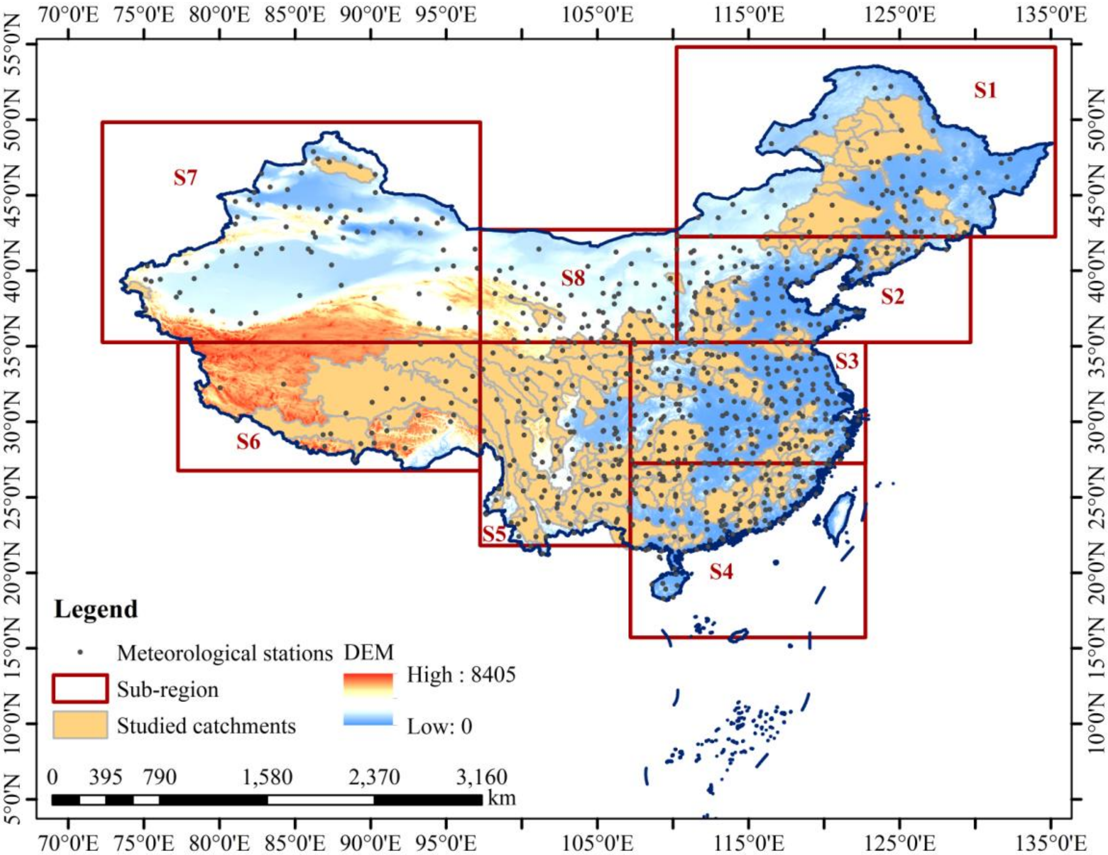

Figure 1: Spatial distribution of meteorological stations over China. The eight geographical sub-regions are plotted (S1-Northeast China, S2-North China, S3-Jiang-Huai Region, S4-South China, S5-Southwest China, S6-Eastern of Tibetan Plateau, S7-Western of Northwest China, and S8-Eastern of Northwest China)**.**

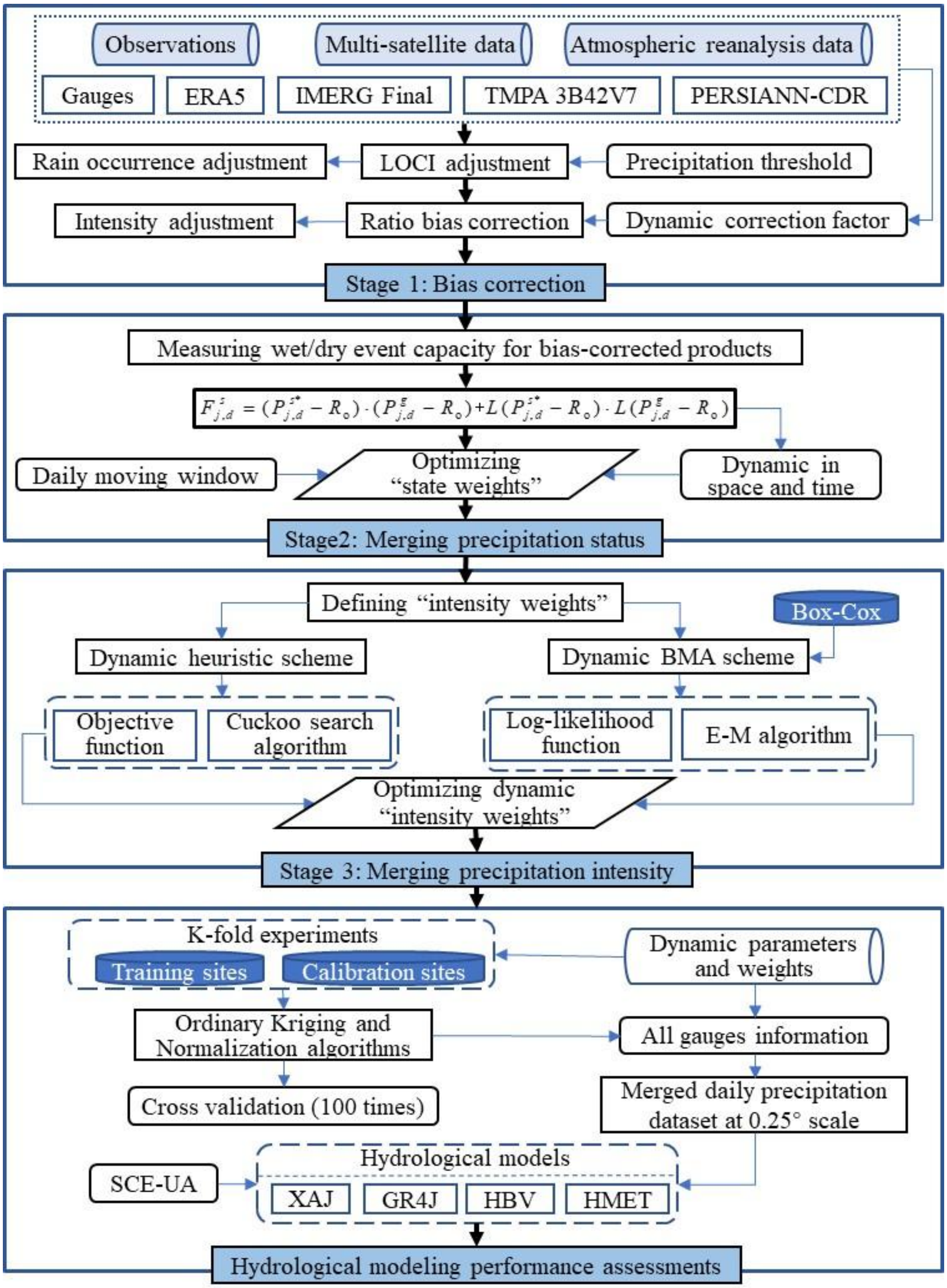

Figure 2: The diagram of the developed three-stage blending framework and its performance validation procedures.

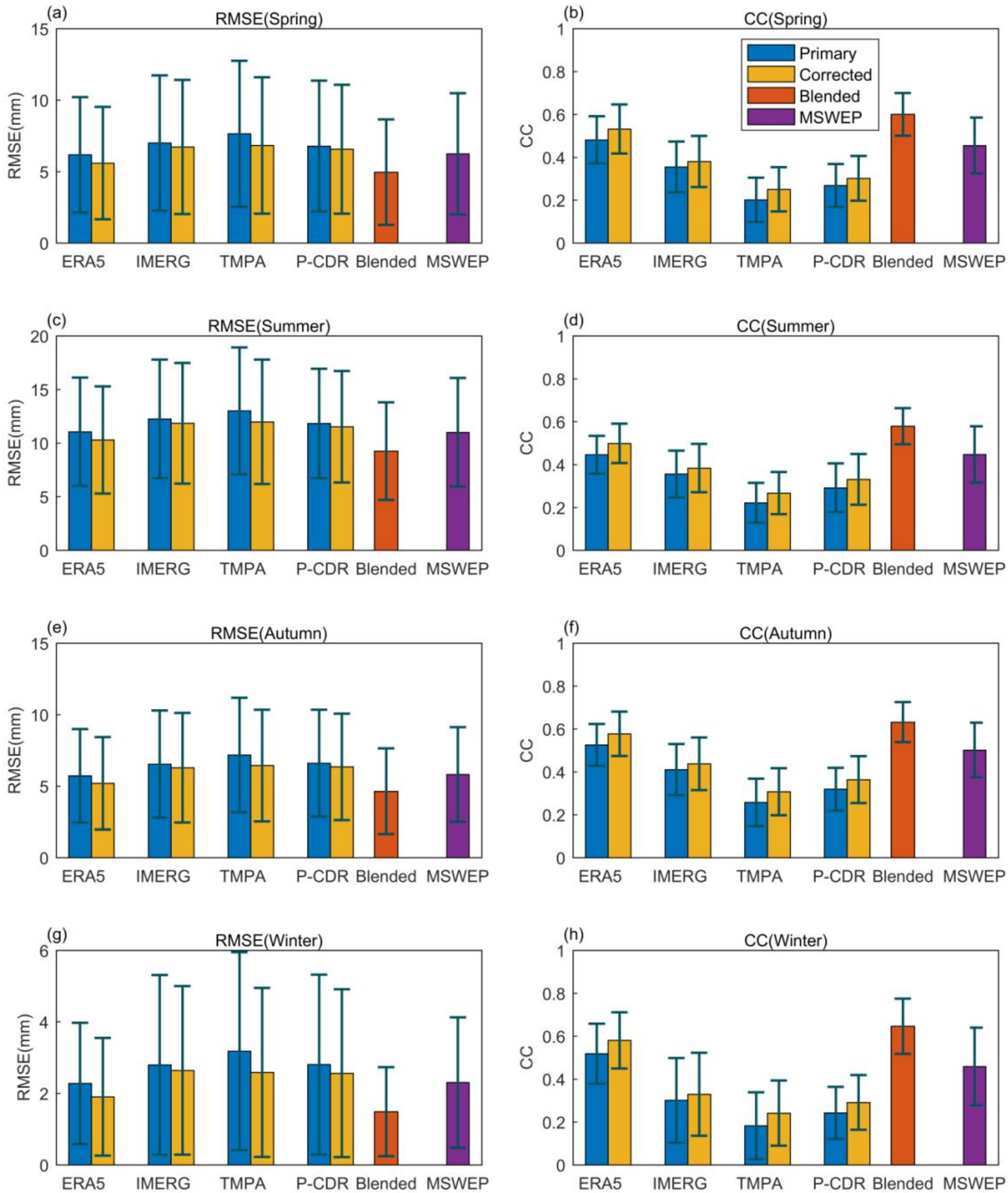

Figure 3: Seasonal average statistical results of RMSE and CC metrics for primary, bias-corrected and three-stage CS blended precipitation products over mainland China. The whiskers denote standard deviation.

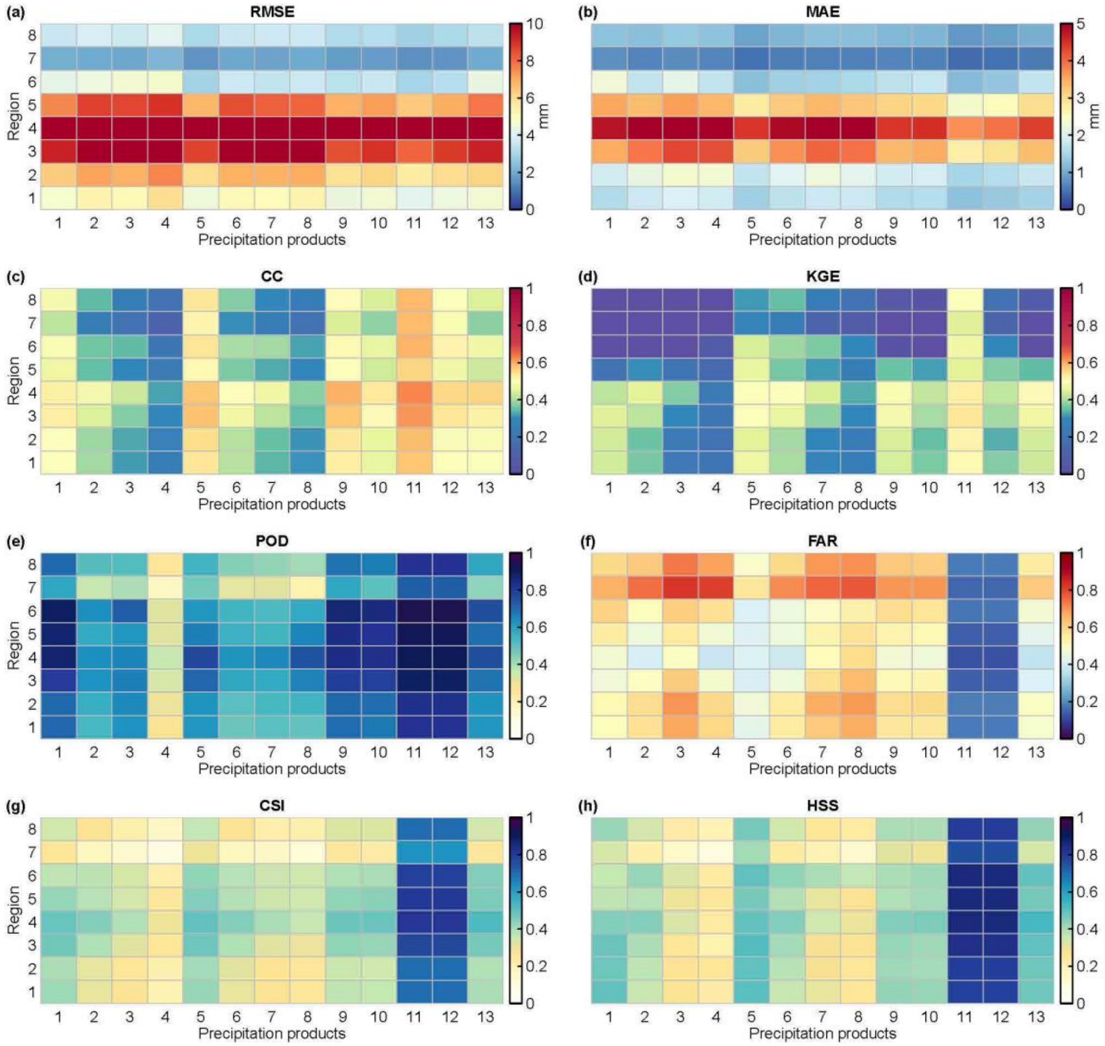

Figure 4: Average statistical metrics of primary, bias-corrected and blended precipitation products in eight sub-regions over China. In the figure, the numbers on the y-axis correspond to eight sub-regions, whereas the numbers on the x-axis correspond to the products identified in Table 2.

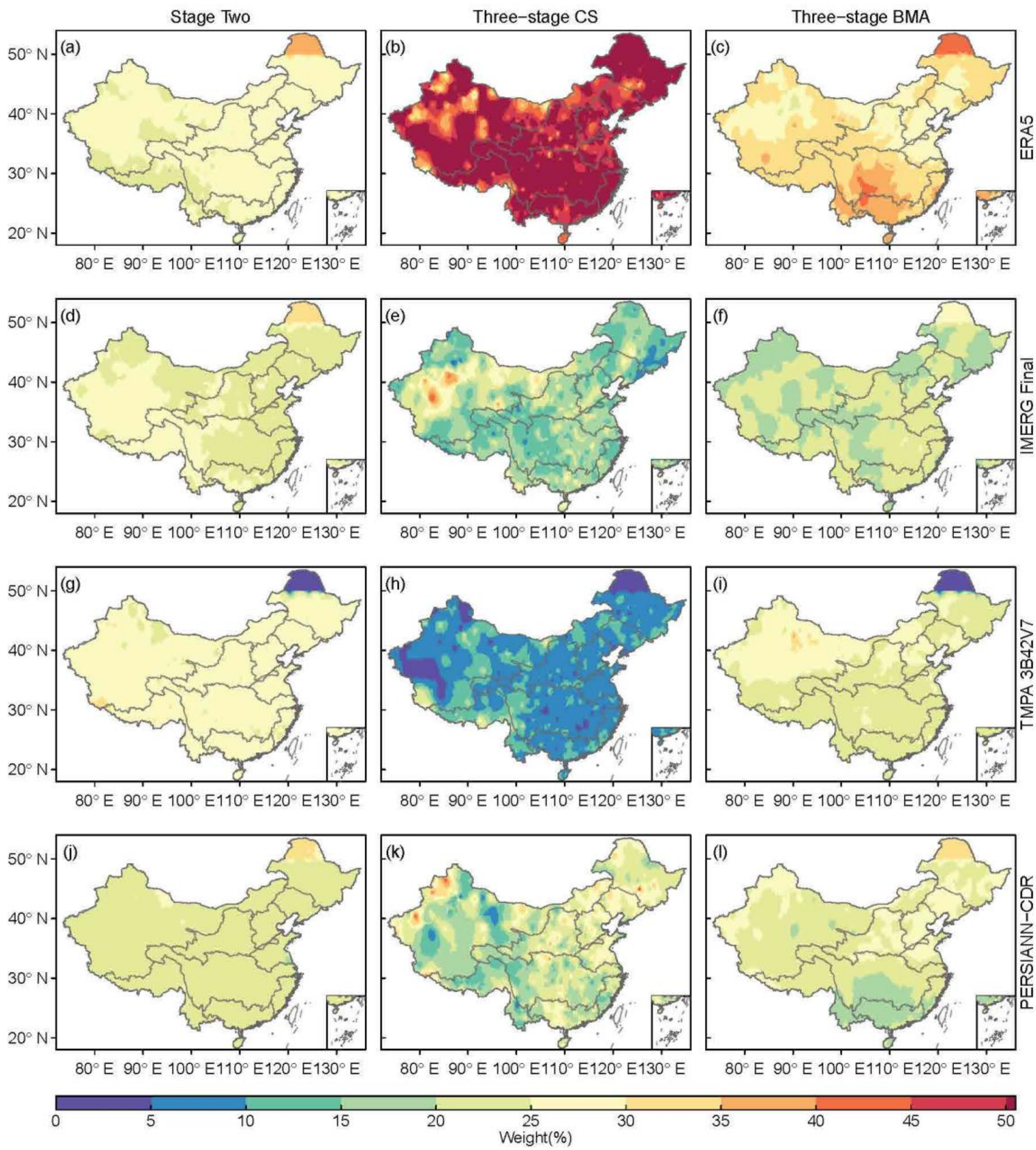

Figure 5: Spatial distributions of average weights of four bias-corrected precipitation members in different blending schemes over mainland China.

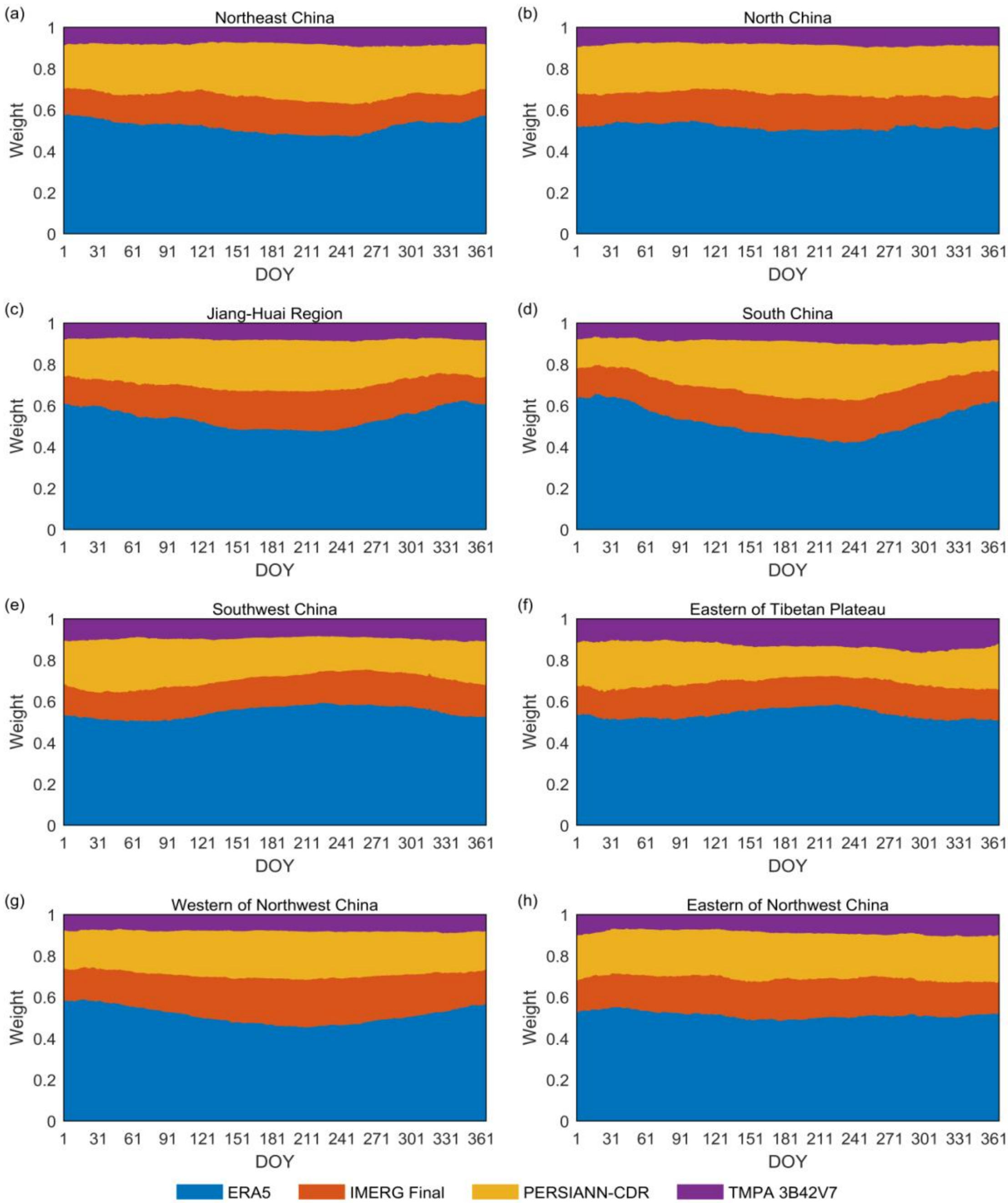

Figure 6: Temporal distribution of average weights in three-stage CS blending scheme for eight sub-regions over mainland China.

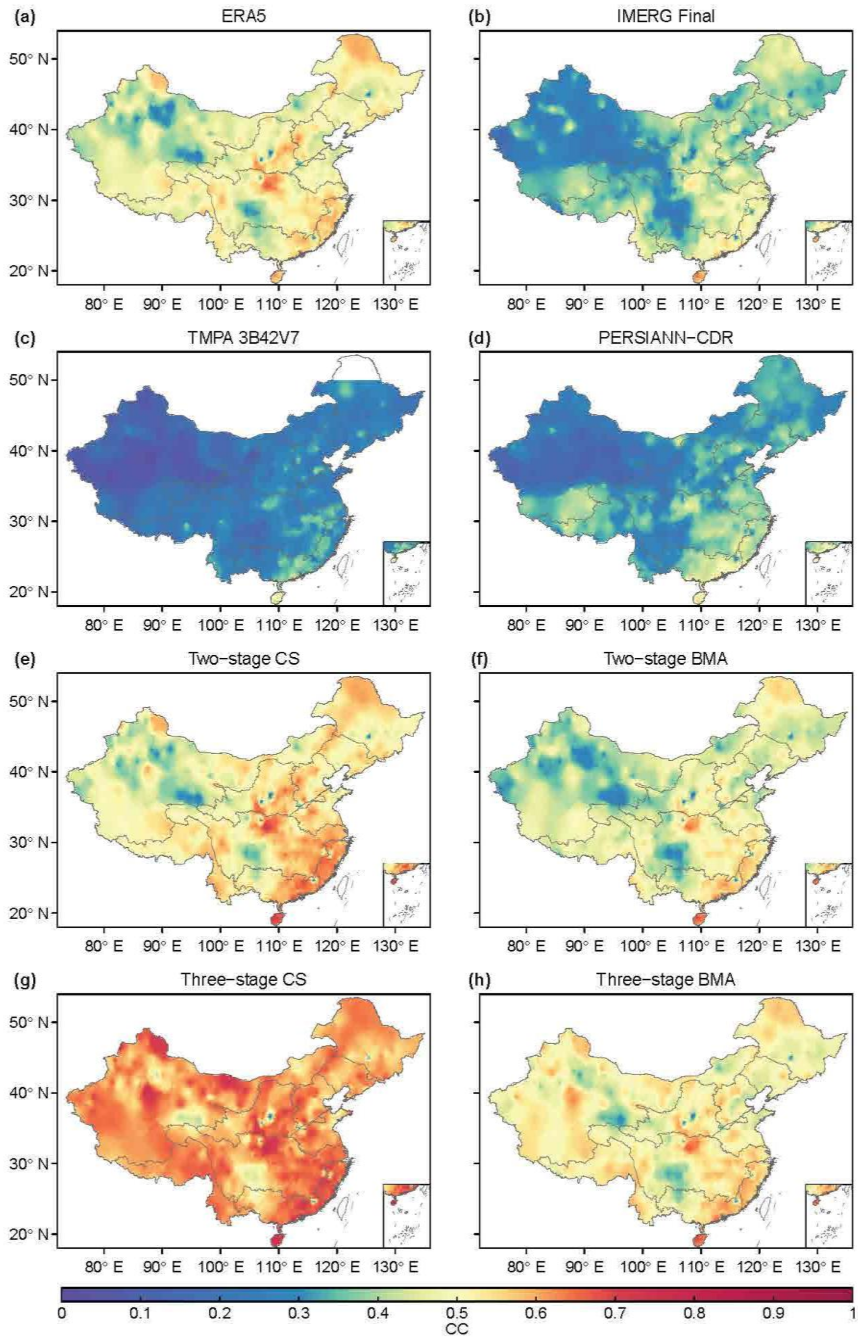

Figure 7: Spatial distributions of statistical metric CC values for primary individuals and four blended ensembles over mainland China.

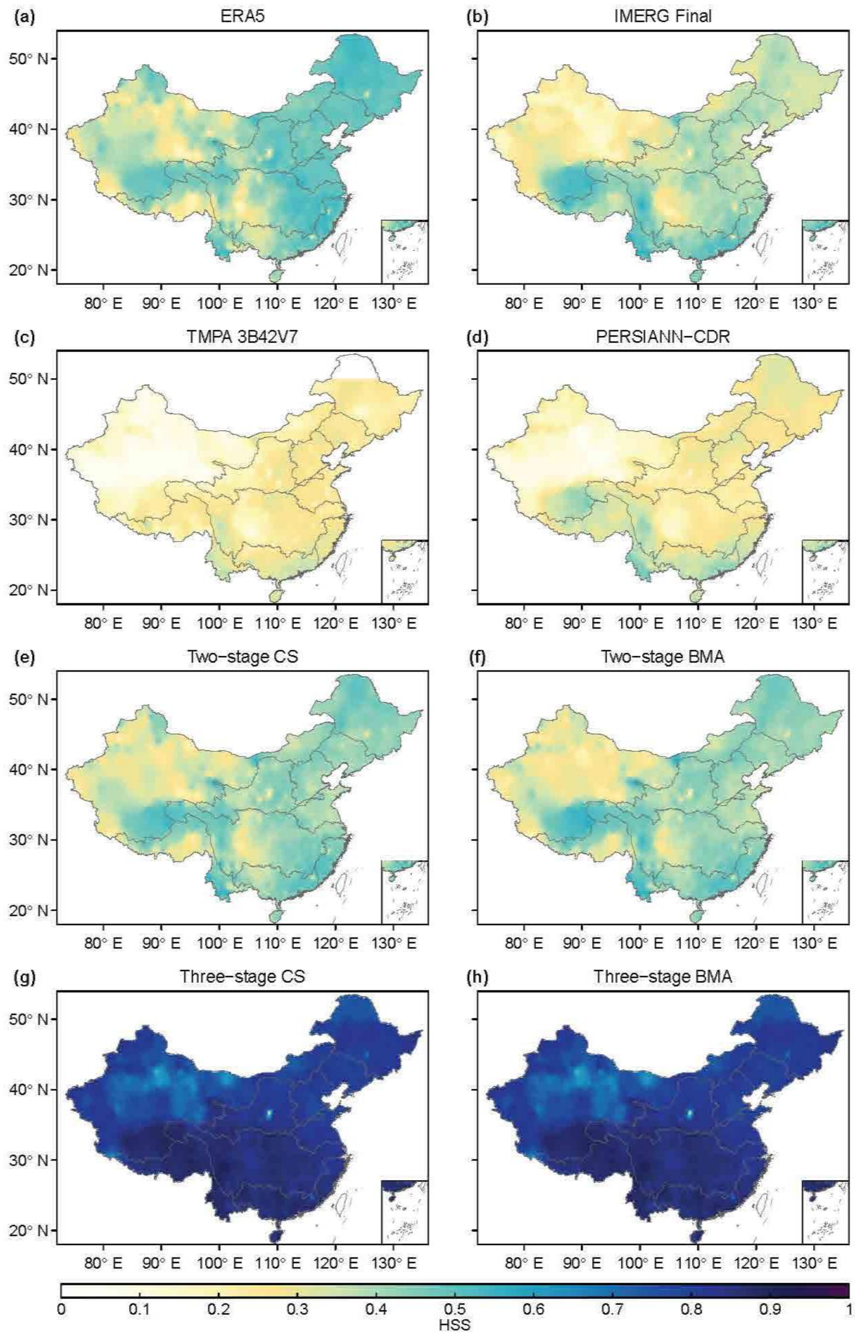

Figure 8: Spatial distributions of statistical metric HSS values for primary individuals and four blended ensembles over mainland China.

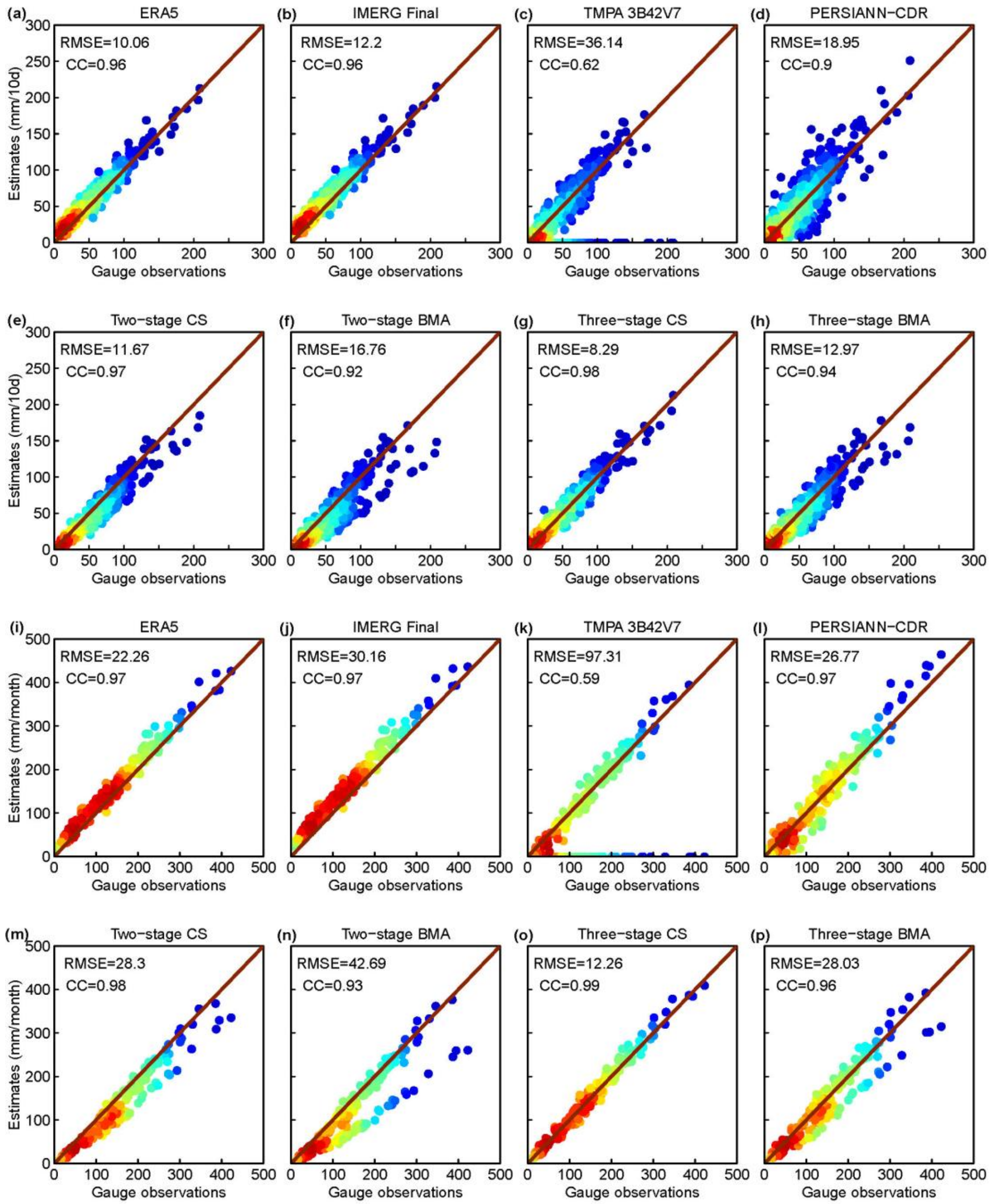

Figure 9: Scatter plots of ten-days and monthly areal precipitation from primary and blended products against gauge-based precipitation in South China.

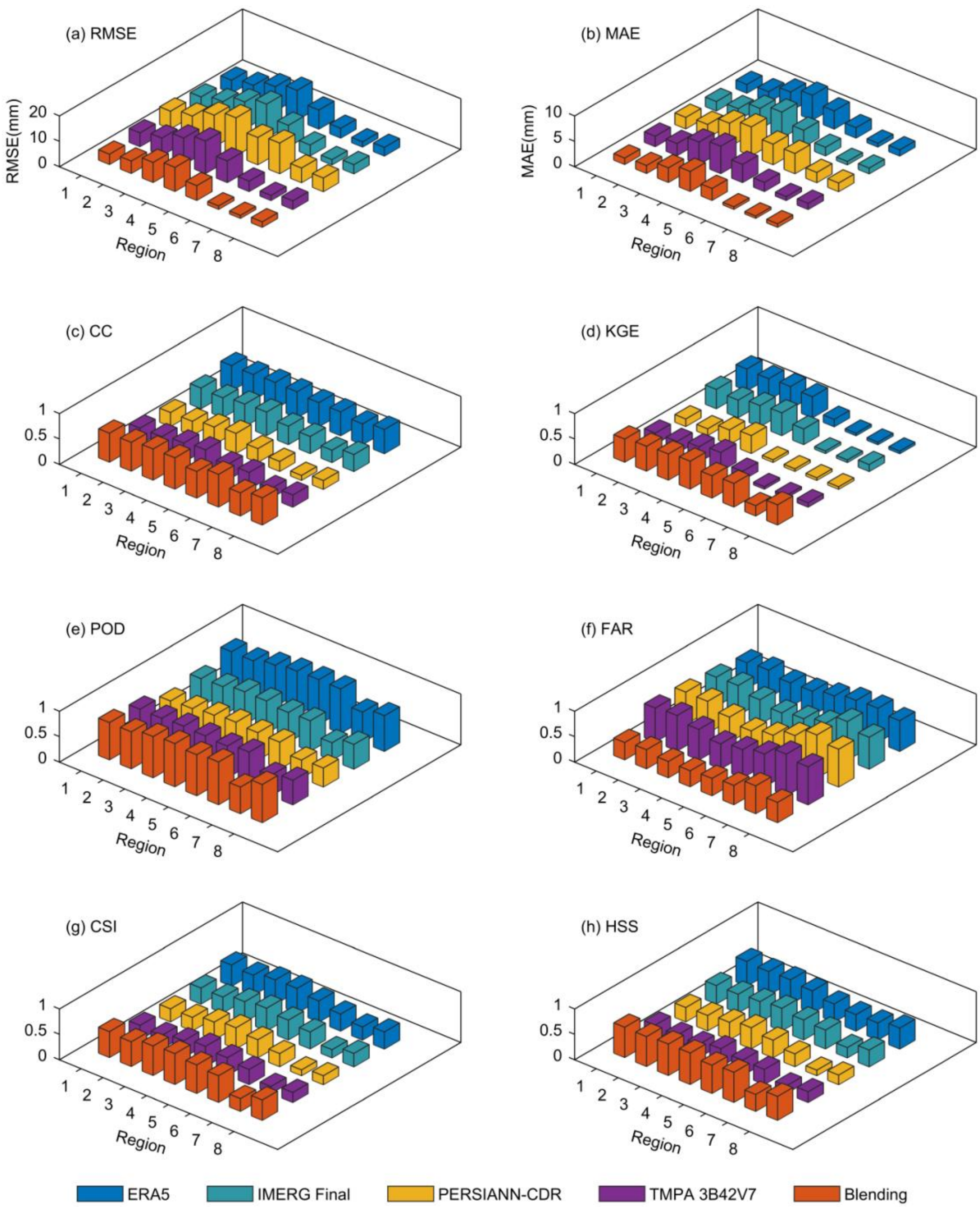

Figure 10: Average statistical metrics of primary and three-stage CS blended precipitation products in random calibration sites over eight sub-regions.

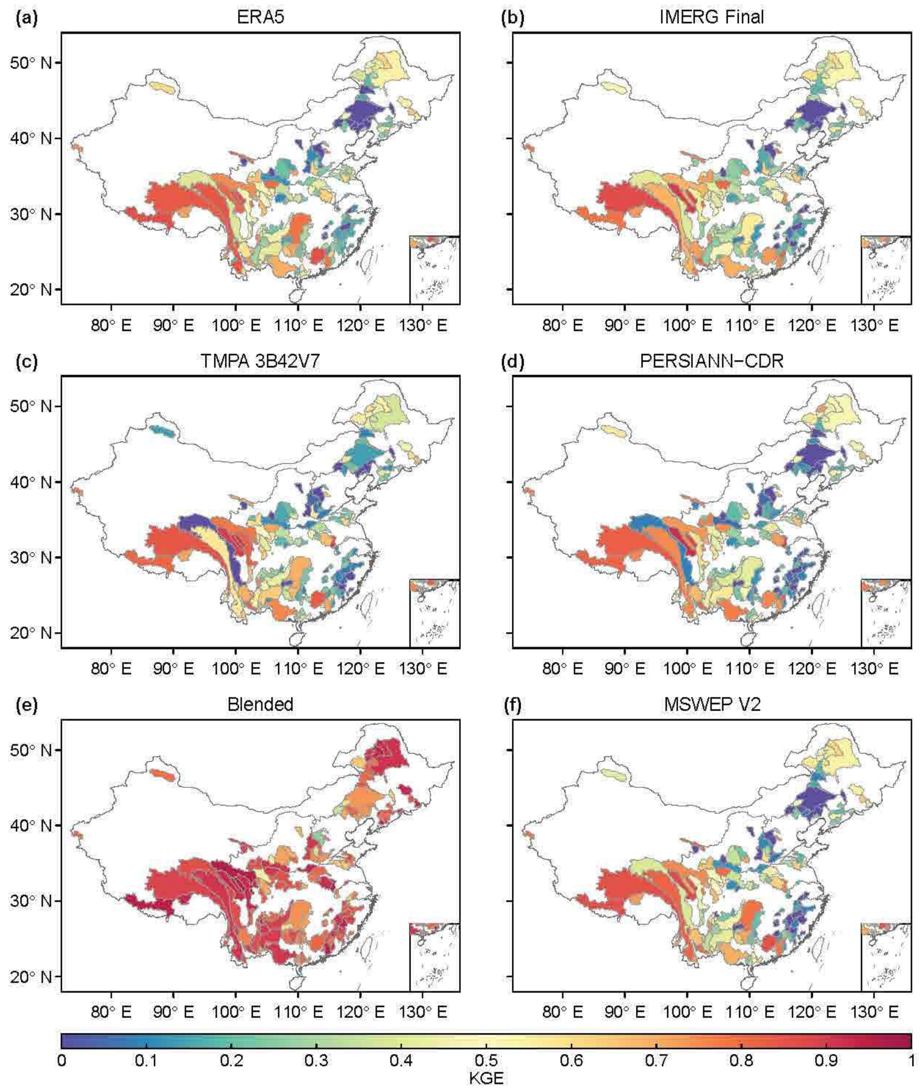

Figure 11: The KGE value of hydrological simulations forcing by primary, three-stage CS blended and MSWEP V2 precipitation datasets during calibration period over 238 catchments. The KGE values correspond to the model that best performed in each catchment.

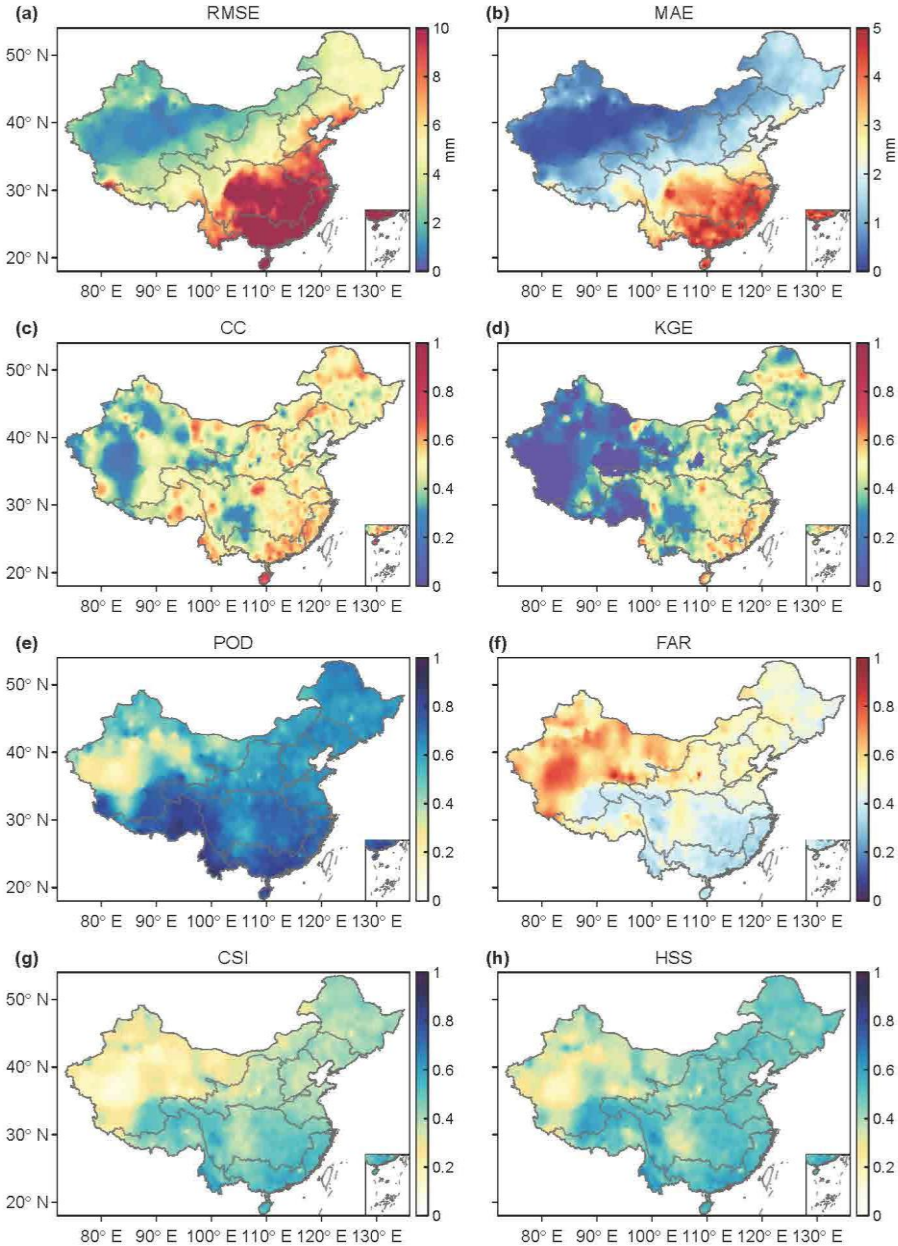

Figure 12: Spatial distributions of eight statistical metric values for MSWEP V2 dataset over mainland China.

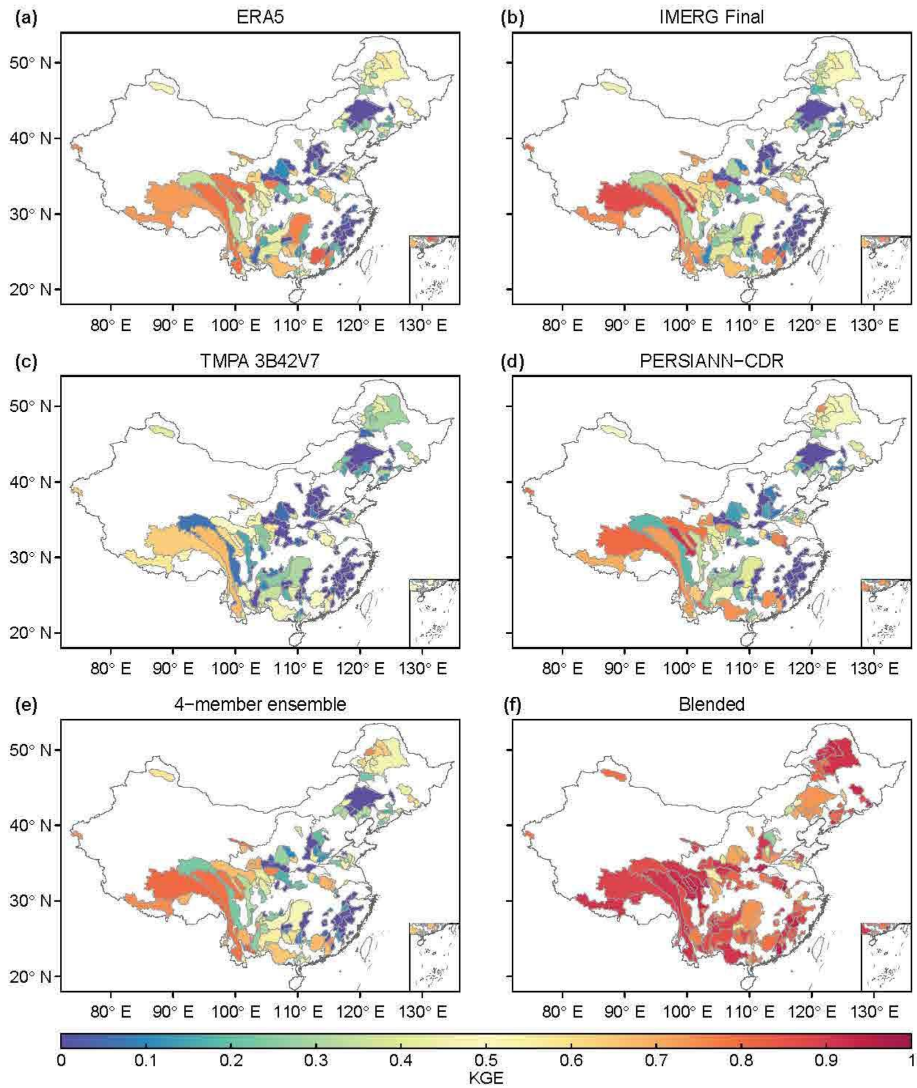

Figure 13: The KGE value of Xinanjiang model when forced by individual precipitation datasets and 4-member ensemble during calibration period over 238 catchments. The "4-member ensemble" denotes that the model is forced by four precipitation datasets and then the simulated streamflow is blended by dynamic BMA method.

***Declaration of Interest Statement**

Declaration of interests

☒ 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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Author Contribution Statement:

J. Yin and S. Guo: Conceptualization; J. Yin, L. Gu and J. Chen: Data curation; J. Yin, D. Liu and Y. Shen: Original draft preparation; Z. Zeng and Y. Shen: Visualization; S. Guo and C.-Y. Xu: Supervision.

Abstract: Satellite-retrieved and atmospheric reanalysis precipitation can bridge the spatiotemporal gaps of in-situ gauging networks, but estimation biases can limit their reliable applications in hydrological monitoring and modelling. To correct precipitation occurrence and intensity simultaneously, this study develops a three-stage blending approach to integrate three multi-satellite precipitation datasets (IMERG Final, TMPA 3B42V7 and PERSIANN-CDR), the ERA5 atmospheric reanalysis product and a gauge dataset within a dynamic framework. Firstly, the systematic biases of the four members were individually corrected by combining the local intensity scaling and ratio bias correction methods. Then, the "state weights" used for determining wet/dry events were optimized by evaluating a score function of the four bias-corrected members. Thirdly, the "intensity weights" were optimized using the cuckoo search (CS) algorithm and the Bayesian Model Averaging (BMA) method, respectively. The three-stage blending approach produced dynamic weights varying both spatially and temporally, and the performance was thoroughly evaluated over mainland China. Results show that the three-stage dynamic scheme performs better than individual datasets and two-stage blending methods in terms of all eight statistical metrics, and the CS algorithm outperforms the BMA method in the third stage. By randomly sampling validation sites using K-fold experiments, the developed algorithm also demonstrates a superior performance in ungauged regions. After interpolating and normalizing blending parameters of all gauges to entire domain using ordinary kriging, a new blended precipitation dataset with a daily 0.25° scale was produced. Four hydrological models are forced by blended and primary

precipitations in 238 catchments over China, further confirming that the developed approach can facilitate hydrological modelling demonstrated by improving the Kling-Gupta efficiency of simulated streamflow by 12-35%.