Received: 9 April 2020 | Revised: 10 July 2020 | Accepted: 24 August 2020 | Published on: 8 September 2020

DOI: 10.1002/joc.6813

RESEARCH ARTICLE

International Journal of Climatology RMetS

Spatiotemporal characteristics and estimates of extreme precipitation in the Yangtze River Basin using GLDAS data

Zeqiang Chen[1,2] | **Yi Zeng**[1,2] | **Gaoyun Shen**[1,2] | **Changjiang Xiao**[1,2] | **Lei Xu**[1,2] | **Nengcheng Chen**[1,2]

[1]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China

[2]Collaborative Innovation Centre of Geospatial Technology, Wuhan, China

Correspondence
Nengcheng Chen, State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan, China.
Email: cnc@whu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 41890822, 41971351, 41771422

Abstract

The Yangtze River Basin has periodically been subject to torrential rains and floods. It is of great significance to characterize the extreme precipitation patterns to learn about frequent flood characteristics in the Yangtze River Basin. Commonly, spatiotemporal characteristics of extreme precipitation was studied by regional frequency analysis method with site data. Spatial sparse site data may cause imprecise divisions of homogeneous regions. In this paper, the spatiotemporal characteristics of extreme precipitation was studied by regional frequency analysis with corrected satellite-based grid precipitation data (Global Land Data Assimilation System, GLDAS) rather than site data. The results show that: (1) The corrected GLDAS daily precipitation data had greatly improved its ability to capture extreme precipitation events in Yangtze River Basin, as the data average accuracy increased from 0.215 before correction to 0.849 after correction. It is feasible to use satellite-based grid precipitation data to replace the site data for the regional frequency analysis of extreme precipitation. (2) The Yangtze River Basin was categorized into seven homogeneous regions for the annual maximum 1-day (RX1DAY) index with an automatic subjective adjustment method. (3) The regional growth curves and quantiles of the Yangtze River Basin were drawn for the return period for 2–100 years. (4) Spatial patterns of extreme daily precipitation series with a return period of 100 years indicated that the precipitation amount increases gradually from the upper to the lower Yangtze River Basin, from the "arid zone" to the "wet zone" and then to the "special wet zone", and the 100-year return level of RX1DAY varied from 30.3 to 301.8 mm. There were three main precipitation centres, the Sichuan Basin, Dongting Lake Basin, and a great triangle area covering the Poyang Lake Basin and the south foot of Dabie Mountain.

KEYWORDS

automatic subjective adjustment method, extreme precipitation, GLDAS, regional frequency analysis, Yangtze River Basin

1 | INTRODUCTION

During the past decades, the Yangtze River Basin has been characterized by vigorous social and economic development and population density. The population and economy are increasingly exposed to frequent flood disasters, which cause huge losses (Gemmer *et al*., 2008). In the process of global warming, many flood events occurred in the Yangtze River Basin (Hlavcova *et al*., 2015), among which the flood caused by extreme precipitation accounts for a large proportion (Wang *et al*., 2013). This phenomenon has raised concerns regarding the increasing frequency and intensity of extreme precipitation (Guo *et al*., 2013, 2016; Gao and Xie, 2016; Lü *et al*., 2018; Wu *et al*., 2018a, 2018b). Therefore, it is important to analyse the intensity, frequency, spatial pattern, and temporal trend of extreme precipitation in the Yangtze River Basin.

The regional frequency analysis method was provided to be suitable, accurate, and robust for the study of extreme precipitation spatiotemporal characteristics (Hosking *et al*., 1985), and the regional frequency analysis method based on L-moments has been extensively applied in many areas (Alvarez *et al*., 2016; Liang *et al*., 2017). However, many studies using precipitation data from meteorological stations for regional frequency analysis (Anli *et al*., 2009; Chen *et al*., 2014b) may result in two problems: one is sparsely distributed stations causing an approximate division with imprecise boundaries (Yang *et al*., 2010); the other is that the method cannot be applied in an area without station coverage. Considering the problems, satellite-based grid precipitation products and reanalysis datasets, such as GLDAS (Bai and Liu, 2018), TRMM (Zhang *et al*., 2015), GCPP (Basheer and Elagib, 2019), CHIRPS (Usman *et al*., 2018), and PERSIANN-CDR (Gao *et al*., 2018), maybe an alternative data source for regional frequency analysis method as their higher spatial resolution and continuous spatial coverage in station-sparse and ungauged areas. Among these data, the GDAL data record is the longest, which is conducive to long-term frequency analysis. Therefore, GDAL data were adopted in this paper.

Although precipitation products can make up for the lack of precipitation data and representativeness to some extent, the accuracy of a single grid or a small time scale cannot meet the requirements of the actual accurate simulation. Precipitation bias on different spatial and temporal scales of nine common satellite-based and reanalysis products had been compared (Tang *et al*., 2020), and the performance of accuracy, distribution, and trends are far from comparable with the ground observation data. The monthly and annual precipitation of GLDAS in the Yangtze River Basin is abnormal (obviously underestimated), and it is recommended to perform error detection and correction during the precipitation simulation process (Zhou *et al*., 2013). Generally, the accuracy of precipitation products at daily scale is not high (Seyyedi *et al*., 2015), the sensitivity to precipitation extreme events is low, and there is a big deviation between precipitation and the actual value (Muhammad *et al*., 2020). The effective evaluation and correction of the precipitation products (Yang *et al*., 2017; Risser *et al*., 2019) is quite necessary for exploring the temporal and spatial distribution of extreme precipitation in a large area with complex terrain. A more refined approach would be to calibrate the GLDAS with the machine learning method (Wang *et al*., 2016) before the regional frequency analysis of extreme precipitation characteristics in the Yangtze River Basin.

The input of regional frequency method changes from site data to satellite-based grid data, which not only changes the type of input data but also brings some challenges to the method. As we know, the regional frequency analysis generally maintains four steps as data screening for the analysis, formation of homogeneous regions for the study area, regional frequency analysis for the target series, and description of spatiotemporal characteristics. In the data screening for the analysis step, Hosking's hypothesis on the stationary and independence of frequency analysis was not fully taken into account in many studies (Yin *et al*., 2016) because of site data observed far from each other. However, the satellite-based grid data are continuous coverage data, failure to test for stability and independence may lead to erroneous conclusions. The inter-site correlation (inter-site independence), which would lead to overly idealized regional analysis results, is most overlooked. The stationarity and dependency of applied data should be focused on to discuss the regional frequency analysis results. In the formation of homogeneous regions for study area step, adjustment of homogeneous regions is time-consuming and energy-consuming (Hosking and Wallis, 1997), because of the manual work, which did not appear to provide a definite improvement in previous studies (Sun *et al*., 2017). The time spent in adjustment will significantly increase with increasing input data from site data to satellite-based grid data and the corresponding number of initial regions. To overcome this problem, an automated adjustment method is needed. The method should efficiently accelerate the identification of homogeneous regions, even in the case of large data input and massive initial clusters.

The objective of the study was to investigate the spatiotemporal characteristics of extreme precipitation and regionalization in the Yangtze River Basin using the regional frequency analysis based on the satellite-based grid precipitation product GLDAS with a new automatic subjective adjustment method. The contributions of the

study are as follows: (1) Using corrected satellite-based grid precipitation products instead of the sparsely distributed site data as the input of the regional frequency analysis, make the regional frequency analysis more accurate in spatial coverage. (2) An automatic subjective adjustment method is proposed to advance the previous method of manual adjustment; it improves the adjustment efficiency and the process is more standardized. (3) Seven homogeneous regions are identified at the basin scale in terms of annual maximum 1-day precipitation index, making the analysis more targeted in determining the temporal and spatial distribution of extreme precipitation in the Yangtze River Basin.

The study and data are described in the next section. The basic employed methods are described in Section 3. The correction of GLDAS dataset, spatiotemporal characteristic results, particularly the frequency of extreme precipitation events, are addressed in Section 4, the summary and conclusions are presented in Section 5.

2 | STUDY AREA AND DATA

2.1 | Study area

The Yangtze River (Figure 1), the longest river in China and supporting the thriving socio-economy of the Yangtze River Basin, has been subject to floods that threaten lives and property for a long time (Xia *et al*., 2001). With an average rainfall as 1,076 mm, extreme precipitation events are the main causes of flood disasters in the Yangtze River Basin. Because of the vast territory, complex topography, and typical monsoonal climate in the Yangtze River Basin, the spatial and temporal distribution of precipitation is extraordinarily uneven (Chen *et al*., 2014a). The frequent flood disasters in the Yangtze River Basin are closely related to the uneven distribution of precipitation variations (Jiang *et al*., 2010). Therefore, the regionalization and quantification with respect to precipitation extremes in this basin are of paramount importance and are worthy of further study.

2.2 | Data

2.2.1 | GLDAS dataset

The daily dataset, GLDAS_CLSM025_D.2.0, is simulated from the Catchment Land Surface Model (CLSM), with a long time span covering from 1948 to 2014 (https://hydro1.gesdisc.eosdis.nasa.gov/data/GLDAS/GLDAS_CLSM025_D.2.0/). CLSM was advanced and developed by the NASA Global Modelling and Assimilation Office (Koster *et al*., 2000). The dataset was forced by Princeton University's Global Meteorological Forcing Dataset. For more information about GLDAS V2.0 and CLSM, please see the GES DISC website (https://disc.gsfc.nasa). Compared with other precipitation products, the GLDAS data set has a longer recording period and vast applications (Li *et al*., 2018; Li *et al*., 2020), which is conducive to the regional frequency analysis of a longer return period (such as 100 years).

The GLDAS_CLSM025_D2.0 dataset was extracted at a basin-scale by using Python with batch processing. According to the 11 extreme precipitation indices proposed by the World Meteorological Organization (WMO) Expert Group on Climate Change Monitoring, Detection and Index (ETCCDMI) (ETCCDI/CRD, 2013; Chen *et al*., 2014b), we chose the annual maximum 1-day precipitation (RX1DAY index) to represent extreme precipitation events. The precipitation variable, total precipitation rate, was multiplied by the number of seconds per day, and then the input precipitation data in mm was obtained.

2.2.2 | Meteorological station data

Daily data set for surface climatological information for China (V3.0) (http://data.cma.cn/data/cdcdetail/dataCode/SURF_CLI_CHN_MUL_DAY_V3.0.html), including meteorological factors such as daily average pressure, maximum pressure, minimum pressure, sunshine duration, average

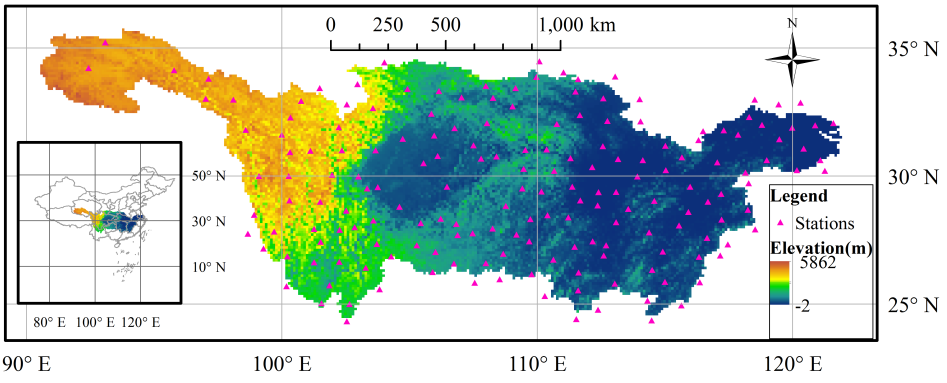

FIGURE 1 The Yangtze River Basin and meteorological stations

temperature, maximum temperature, average wind speed, maximum wind speed and direction, extreme wind speed and direction, average relative temperature, minimum relative humidity, and precipitation. The data range is from 1951 to the present.

The meteorological data of the stations have been processed by the data provider, such as climatic limit value (allowable value) check, regular check, consistency check of extreme value and average value, spatiotemporal consistency check, as well as manual verification and correction. The actual and correct rate of the data is close to 100%.

3 | METHODOLOGY

The method with five main parts that characterizes the spatial pattern of precipitation extremes in the Yangtze River Basin is shown in Figure 2. The five parts are (1) Correction and accuracy assessment: correcting the GLDAS daily precipitation data and provide high-precision and high-resolution precipitation input, (2) Data screening: screening available data for RX1DAY series, (3) Formation of homogeneous regions: identifying the homogeneous regions for the Yangtze River Basin, (4) Regional Frequency analysis: performing regional frequency analysis for each homogeneous region, and (5) Spatiotemporal characterization: executing the ordinary kriging interpolation method (Bregt *et al*., 1991) in ArcMap (Silvestre *et al*., 2016) to visually describe the extreme precipitation spatial pattern under the return period of 100 years. A detailed description of the main methods and procedures used in the first four parts is given in the following sections.

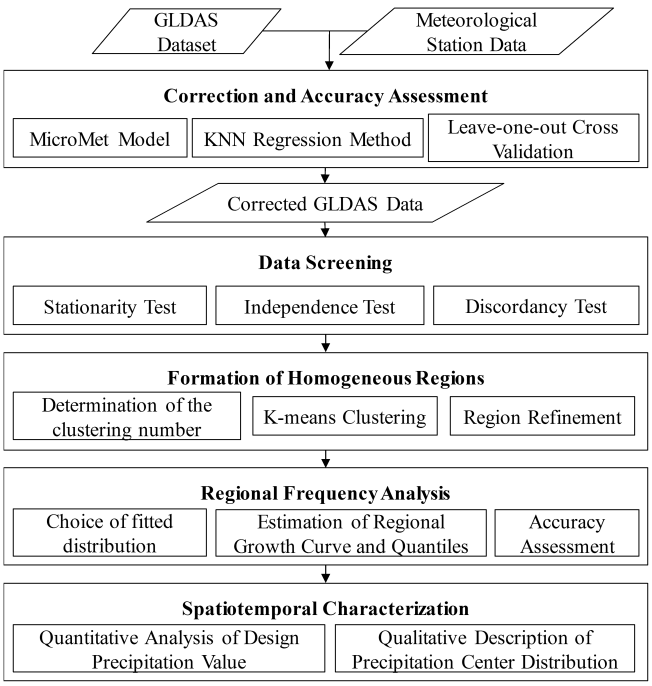

FIGURE 2 The technological process that characterize the spatial pattern of precipitation extremes

3.1 | GLDAS dataset correction and accuracy assessment

In the production process of satellite-based precipitation products, due to the introduction of some systematic errors in observation methods and inversion algorithms, their accuracy in a single grid range and a small time scale cannot meet the analysis requirements. Based on the reference surfaces generated by the meteorological data of the station with MicroMet model, this paper adopts a machine learning method, KKN regression model, to correct the GLDAS daily data.

3.1.1 | MicroMet model

MicroMet can provide high-resolution meteorology-driven forcing data, and interpolate the atmospheric-driven factors by the combination of mathematical method, empirical formula, and physical formula. Using the data from the ground meteorological station, MicroMet model can generate filed data for the common meteorological elements: air temperature, relative humidity, wind speed, wind direction and precipitation, etc. In order to create a distributed atmospheric field, the general steps are: (1) interpolate on the meteorological elements, (2) use the known temperature-altitude, wind-topography, and humidity-cloud relationships to subsequently correct the interpolation field (Liston and Elder, 2006).

The station precipitation data and elevation information are used to produce the grid precipitation (P_0) by the distance weighted interpolation method. The precipitation adjustment coefficient (lapse rate, χ) is calculated according to the precipitation observation data as follows,

$$\chi = \frac{p_i - p_j}{p_i + p_j} / \left(Z_i - Z_j\right) \tag{1}$$

The interpolation results of each grid (P) are adjusted with reference to the precipitation-elevation equation (as follows) and precipitation lapse rate χ.

$$P = P_0 \left[\frac{1+\chi(z-z_0)}{1-\chi(z-z_0)}\right] \tag{2}$$

Adjust the temperature (T_{air}) of all stations to the same altitude (e.g., sea level). The distance weighted

interpolation method is used to interpolate the adjusted station temperature into the grid. The terrain data and the temperature lapse rate (Kunkel, 1989) are used to adjust the grid temperature back to the actual altitude.

The relative humidity needs to be analysed on the basis of air temperature. Convert the relative humidity (rh) of the station into dew temperature (Td) as Equation 2 (Kunkel, 1989). Adjust the dew temperature of all stations to the same altitude (such as sea level) by using the dew lapse rate. The dew temperature of the station is interpolated to the sea level grid. The dew lapse rate is used to get the grid to the corresponding value of the actual altitude. Calculate the dew point temperature back to relative humidity.

$$es=\frac{A\sqrt{B\left(T_{air}-T_f\right)}}{C+T_{air}-T_f}, e=\frac{es\times max(10.0,rh)}{100.0}, Td=\frac{C\times log\left(\frac{e}{A}\right)}{B-log\left(\frac{e}{A}\right)}+T_f \quad (3)$$

The default parameters in the empirical formula are $A = 6.1121 \times 100.0$, B = 17.502, and C = 240.97. And the temperature is $T_f = 273.16$.

Finally, the above meteorological factors are interpolated as reference surfaces to correct the GLDAS daily precipitation.

3.1.2 | KNN regression correction

K nearest neighbour (KNN) algorithm is a mature and simple supervised machine learning model (Mehdizadeh, 2020). In this paper, it is mainly applied to the process of precipitation correction as a non-parametric regression method suitable for uncertain and nonlinear dynamic systems. KNN model has better performance when dealing with samples with the uneven spatial distribution.

Because the accuracy of GLDAS daily precipitation data is poor (underestimated to a large extent), the correction of daily precipitation needs to be further carried out step by step according to the precipitation level. The range of daily rainfall value in the "Precipitation intensity classification standard (inland part)" (http://www.gov.cn/ztzl/2008tffy/content_1113935.htm) issued by China Meteorological Administration in 2006 was taken as the classification standard, and the daily rainfall data to be corrected were divided into six levels, that is, light rain (<10 mm), moderate rain (10 – 25 mm), heavy rain (25 – 50 mm), torrential rain (50 – 100 mm), downpour (100 – 250 mm), and extremely heavy rain (>250 mm).

The given training sample set is the data pair set containing the input vector and the output value, $\{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$. The input vector of the sample, $x_i = (x_{i1}, x_{i2}, \cdots, x_{im})$ is designed for the research needs, y_i is the output value. Then the d_i distance between the current sample points is calculated according to the distance measure (e.g., Euclidean distance). Then the nearest neighbour elements are obtained by sorting the distances in ascending order. The output values of these nearest neighbours are recorded as $y_i(x)$. According to the leave-one-out cross-validation results, the optimal nearest neighbour K is determined, and the output value of the final target sample point can be obtained from the mean value of the output value of the KNN as follows,

$$\hat{y}=\frac{1}{N}\sum_{i=1}^{N} y_i(x) \quad (4)$$

Here, eight characteristic variables that have a great impact on precipitation, longitude, latitude, elevation, slope, aspect, relative humidity, air temperature, and the GLDAS precipitation to be corrected are selected as the input vector, and the output value is the actual daily precipitation. The R package for weighted KNN regression, kknn v1.3.1 (https://CRAN.R-project.org/package=kknn), is used to predict, or rather, to correct GLDAS precipitation.

3.2 | Data screening

Data screening plays a decisive role in the regional frequency analysis of extreme precipitation series. In order to satisfy the requirement of regional frequency analysis for data stationarity and independence, the extreme precipitation series should be tested and screened. It is necessary to avoid grids with significant trends and correlations and outliers with obvious inconsistencies in the sequence.

3.2.1 | Tests for stationarity and independency

The extreme precipitation index, RX1DAY is selected in this paper, data screening should be performed for each grids. Two types of screenings were considered: a test for trends and two tests for correlation, which ensured stationary and independent time series for regional frequency analysis.

The modified Mann-Kendall method (MMK) was used to detect the stationarity of the time series. It performs a two-tailed Mann-Kendall test (Hirsch *et al*., 1982) that had been modified to account for autocorrelation on the time series at the 5% significance level.

Autocorrelation for the data value at time lag-1 was computed at each station to assess the independence of the sample data (Wright *et al*., 1988). The absolute value

of the autocorrelation coefficients of lag-1 ($|Lag\_1|$) for a time series consisting of n observations was calculated. Observations can be accepted as being independent from each other when $|Lag\_1|$ is not greater than the critical value, $1.96/\sqrt{n}$, corresponding to the 5% significance level (Yang *et al*., 2010).

The grids that passed the data screening are to be used in the regional frequency analysis, whereas the other grids with significant trends and autocorrelation were excluded from the dataset, and at-site frequency analysis is performed to obtain the complete characteristics of the whole Yangtze River Basin.

3.2.2 | L-moment theory

The regional frequency analysis method in this paper is based on L-moments (Hosking *et al*., 1985), the so-called index flood method. L-moments are more convenient than other conventional moments because they are more easily interpretable as measures of distributional shape, which can be obtained by considering linear combinations of the observations in a sample of data that has been arranged in ascending order (Hosking and Wallis, 1997). Denoted by $X_{k:n}$, the kth smallest observation from an ascending sample of size n, and the L-moment of a probability distribution are defined as follows:

$$\lambda_r = r^{-1}\sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}E(X_{r-k:r}) \tag{5}$$

Thus, λ_1 is the mean of the distribution, λ_2 is the L-scale, τ is the L-CV, τ_3 is the L-skewness, and τ_4 is the L-kurtosis. The sample L-moment ratios, that is, natural estimators of population L-moments, are defined by $t_r = l_r/l_2$, and the sample L-CV is defined by $t = l_2/l_1$.

3.2.3 | Discordancy measure

The discordancy measure was used for identifying unusual sites in a region, that is, the sites where the at-site sample L-moments are evidently different from those of the other sites in the region. Let $u_i = [t^{(i)}\ t_3^{(i)}\ t_4^{(i)}]^T$ be a transposition vector containing the t, t_3, and t_4 values for the site i in a data set containing N sites. Let $\bar{u} = N^{-1}\sum_{i=1}^{N} u_i$ be the (unweighted) dataset average. So the discordancy measure for the site i is as follows:

$$D_i = \frac{1}{3}N(u_i-\bar{u})^T A^{-1}(u_i-\bar{u}) \tag{6}$$

where A is defined as $\sum_{i=1}^{N}(u_i-\bar{u})(u_i-\bar{u})^T$.

Large values of D_i imply sites that are discordant from the data set and even likely to be data errors. Any site with $D_i \geq 3$ or $D_i > (\mathrm{N}-1)/3$ should be regarded as discordant (Hosking and Wallis, 1997).

3.3 | Formation of homogeneous regions

The identification of precipitation consistent region in the Yangtze River Basin is the premise of regional frequency analysis. In this section, on the basis of the corrected GLDAS daily precipitation data, the appropriate precipitation climate characterization factors will be selected to cluster the grid precipitation in the Yangtze River Basin. Then, the automatic subjective adjustment method proposed in this paper can reduce the heterogeneity of clustering results, so that the Yangtze River Basin can be divided into several spatially continuous and physically reasonable precipitation homogenous regions.

3.3.1 | K-means clustering

A cluster analysis of site characteristics is a standard and practical method for forming regions from large data sets for regional frequency analysis (Goyal and Gupta, 2014). There is a need to seek a proper number of clusters for an appropriate clustering, which is a good basis for forming regions (Hosking and Wallis, 1997). Using the Davies-Bouldin index, we can obtain the optimal clustering number for k-means clustering (Maulik and Bandyopadhyay, 2002).

The k-means clustering approach takes a dataset X containing N objects as input and a parameter K specifying the cluster number. The output is a set of K cluster centroids (C) and X labels used to assign each object in X to a cluster. The objects in the cluster are as similar as possible and the objects from different clusters are as different as possible. The Davies-Bouldin measure describes the ratio of the sum of the intra-cluster divergence (W) to the distances between cluster centroids:

$$DB = \frac{1}{K}\sum_{i=1}^{K}\max_{j=1\sim K, j\neq i}\left(\frac{W_i+W_j}{C_{ij}}\right) \tag{7}$$

The smaller the DB, the lower the similarity between the clusters and the better the clustering result. K-means clustering is performed with the optimal K and five site characteristics to cluster the grids in the Yangtze River Basin. The five indicators for defining the precipitation

climate here were longitude, latitude, elevation, mean annual precipitation, and the deviation of annual precipitation. Cluster analysis is sensitive to large-scale differences among those variables. The variables are therefore scaled such that their ranges are comparable (Gnanadesikan *et al*., 2007).

The initial obtained clusters are not necessarily the final regions, which need appropriate adjustments to form homogeneous regions (Hosking and Wallis, 1993).

3.3.2 | Heterogeneity measure

A heterogeneity measure was proposed to estimate the heterogeneity in a set of sites and to evaluate whether these sites could be reasonably regarded as a homogeneous region (Hosking and Wallis, 1993). Assuming that there are N sites in the region, with a record length n_i and sample L-moment ratios $t^{(i)}$, $t_3^{(i)}$, and $t_4^{(i)}$ of site i. The regional average L-CV, L-skewness, and L-kurtosis are abbreviated as t^R, t_3^R, and t_4^R, respectively. All ratios are weighted as one because the record length of the grids in the Yangtze River Basin is identical. The dispersion of the at-site sample L-CVs (commonly used) can be calculated as,

$$V = \sqrt{\frac{\sum_{i=1}^{N} n_i (t^i - t^R)^2}{\sum_{i=1}^{N} n_i}} \tag{8}$$

where μ_V and σ_V represent the mean and standard deviation values of derived form 1,000 (large enough for reliability) simulated regions. Therefore, the heterogeneity of the region is measured as follows:

$$H = \frac{V - \mu_V}{\sigma_V} \tag{9}$$

When $H < 1$, the region is "acceptably homogeneous". When $1 \le H < 2$, the region is "possibly heterogeneous". When $H \ge 2$, the region is "definitely heterogeneous" (Hosking and Wallis, 1997). Please refer to Hosking and Wallis (Hosking and Wallis, 1993) for more information about definitions of statistics based on different L-moments.

3.3.3 | Automatic subjective adjustment method

An automatic subjective adjustment method was proposed in this paper. At-site statistics are available for use as the basis for a test of final region homogeneity. The method is a more rapid and efficient iterative process with two criteria of discordancy measure and heterogeneity measure (Hosking and Wallis, 1997). The iteration ends when a final set of acceptable homogeneous regions is obtained.

Hosking presented several types of adjustments to reduce the region heterogeneity, such as moving a site or a few sites from one region to another, deleting a site or a few sites from the data set, subdividing the region, etc. (Hosking and Wallis, 1997). These principles were emphasized and embodied to design automatic subjective adjustment. Figure 3 shows the general automatic subjective adjustment method processes. More detailed rules and a primary algorithm are shown in Appendix A. This method can easily adjust the initial clusters to obtain the final regions.

3.4 | Regional frequency analysis

For the extreme precipitation series, the appropriate frequency distribution function is chosen for all homogeneous regions in the Yangtze River Basin, and then the regional frequency analysis is performed. Considering the heterogeneity and correlation of the actual data, the reliability of the frequency analysis results in each return period is evaluated. Finally, the temporal and spatial characteristics of extreme precipitation in the Yangtze River Basin are revealed quantitatively and qualitatively through the spatial distribution of extreme precipitation in the 100 year return period.

3.4.1 | Choice of the fitted distribution

After the homogeneous regions have been determined, the choice of appropriate frequency distributions should be accordingly selected. An alternative approach that the goodness of fit test of a candidate distribution was proposed to detect whether the regional average L-moments were consistent with the parameters of the fitted distribution (Hosking and Wallis, 1993). This goodness-of-fit measure used a Z^{DIST} statistic as follows:

$$Z^{DIST} = \frac{t_4{}^R - \tau_4{}^{DIST} - B_4}{\sigma_4} \tag{10}$$

where $\tau_4{}^{DIST}$ is the L-kurtosis of the fitted distribution, B_4 and σ_4 are the bias and the standard deviation, respectively. One can declare the fit to be adequate if it is sufficiently near zero, with an acceptable criterion being $|Z^{DIST}| \le 1.64$ (Hosking and Wallis, 1997).

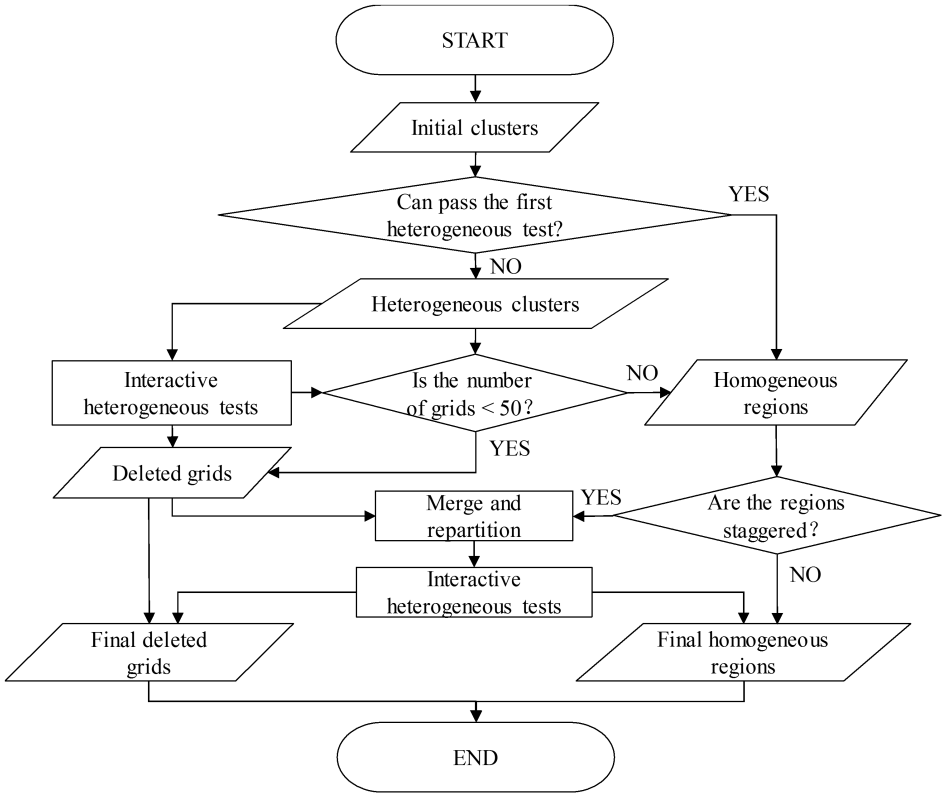

FIGURE 3 Flowchart for automatic subjective adjustment method

For finding a candidate for describing the data, it is necessary to compare the goodness of fit of several commonly used three-parameter distributions, such as, generalized logistic (GLO), generalized extreme-value (GEV), generalized normal (GNO), Pearson type III (PE3), and generalized Pareto (GPA). Particularly when no three-parameter distribution is suitable, a Wakeby distribution with five parameters is used (Hosking and Wallis, 1997). The algorithm used to choose a distribution is shown in Figure 4.

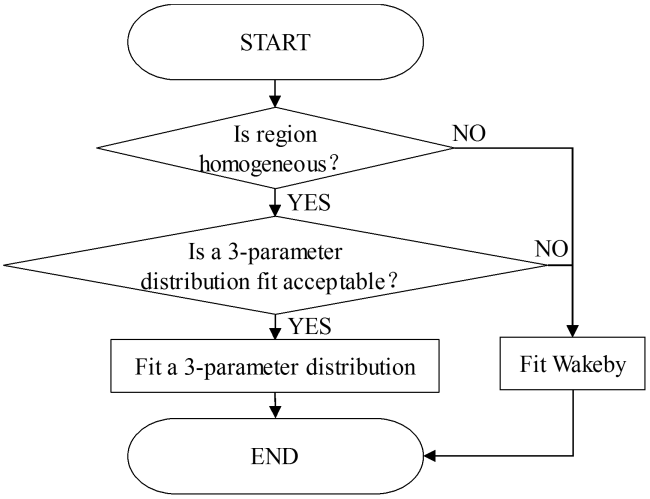

FIGURE 4 Algorithm for choosing a distribution for a region

3.4.2 | Estimation of regional growth curve and quantiles

Sites at which data are available for regional frequency analysis have been assigned to the homogeneous regions, meaning that the frequency at the sites in a region is identical apart from a scale factor (Hosking and Wallis, 1997). Suppose that data are available at N sites, with the site i having a sample size n_i and observed data $P_{ij}, j = 1, \cdots, n_i$. Let $P_i(F)$, $0 < F < 1$ be the quantile function of frequency at the site i. Then,

$$P_i(F) = \mu_i p(F), i = 1, \cdots, N \tag{11}$$

Here, μ_i is the site-specific scaling factor, which is usually estimated by $\hat{\mu}_i = \hat{P}_i$, the expected value of frequency distribution at the corresponding site . The remaining factor $p(F)$ is the regional growth curve, which is a dimensionless quantile that is common to each site and can be computed by $p_{ij} = \frac{P_{ij}}{\hat{\mu}_i}$.

In this approach (Hosking and Wallis, 1997), the L-moments of interest ($\hat{\theta}_k^{(i)}$) are separately estimated at each site, the regional L-moments being an estimate of the at-site L-moments, weighted equally. The estimated regional growth curve $\hat{p}(F) = q\left(F; \hat{\theta}_1^R, \cdots \hat{\theta}_p^R\right)$ can be derived by plugging these at-site estimates into $p(F)$. The

quantile estimates at the site i can be obtained by using the estimates of μ_i and $p(F)$ (Hosking and Wallis, 1997).

3.4.3 | Accessing the accuracy of estimates

The results obtained through statistical analysis are essentially uncertain, and to make the results as useful as possible, some assessment of the uncertainty should be completed (Hosking and Wallis, 1997). Inter-site dependence increases the variability of estimates, although it has little effect on the bias of the estimates.

Monte Carlo simulation is conducted to assess the accuracy of estimations by their root mean square error (RMSE). The simulated region should keep the same conditions as the actual region, such as the number of sites, the record length of each site, the regional average L-moment ratio, and inter-site correlation (Hosking and Wallis, 1997).

Assuming P_i is the precipitation of the grid i, r_{ij} respects the sample correlation between site i and j, then the regional sample correlation ρ can be estimated by the mean value of r_{ij}.

$$r_{ij}=\frac{(P_i-\bar{P}_i)(P_j-\bar{P}_j)}{\sqrt{(P_i-\bar{P}_i)^2(P_j-\bar{P}_j)^2}} \tag{12}$$

According to Hosking' procedure (Hosking and Wallis, 1997), at the *mth* repetition, the quantile estimate at the site i for the specific nonexceedance probability F is $\hat{P}_i^{[m]}(F)$. The relative error of the estimate can be squared and averaged on the basis of all M replicates to estimate the relative RMSE. The approximated relative RMSE at the site i for large M is as follows:

$$R_i(F)=\sqrt{M^{-1}\sum_{m=1}^{M}\left\{\frac{\hat{P}_i^{[m]}(F)-P_i(F)}{P_i(F)}\right\}^2} \tag{13}$$

The regional average relative RMSE of the estimated quantile provides a summary of the estimated quantile accuracy for all sites in the region.

For a specific nonexceedance probability, it may be found that 99% of the distribution lies within the interval $[L_{0.01}(F), U_{.01}(\mathrm{F})]$, where $L_{0.01}(F)$ and $U_{.01}(\mathrm{F})$ are the values at which 99% of the simulated values to the actual value ratio. Therefore, the 99% error bounds are as follows:

$$\frac{\hat{P}(F)}{U_{.01}(F)}\le P(F)\le\frac{\hat{P}(F)}{L_{.01}(F)} \tag{14}$$

The details for the estimation and assessment procedures can be found in Hosking and Wallis (Hosking and Wallis, 1997).

4 | RESULTS AND DISCUSSION

4.1 | Correction of GLDAS dataset

According to the meteorological data of 177 stations (Figure 1) in the 50 km buffer zone around the Yangtze River Basin from 1970 to 2015, the precipitation lapse rate, air temperature lapse rate, and dew temperature lapse rate were obtained, and various meteorological reference surfaces for GLDAS precipitation correction were obtained by the MicroMet method.

To illustrate the correction process of GLDAS daily precipitation, here the flood of Yangtze River caused by extreme precipitation in 1998 was used as an example. According to the station observation record, the maximum daily precipitation of 1998 happened on July 22. The air temperature reference surface and dew temperature reference surface on July 22, 1998, are shown in Figure 5a,b.

There were three levels of precipitation data sets at 177 stations in the entire Yangtze River Basin. Table 1 showed the optimal K (bestK) and corresponding RMSE of the KNN models of the three datasets for that day by leave-one-out cross-validation.

With the KNN regression models of each day from January 1, 1970, to December 31, 2014, the daily GLDAS precipitation was finally corrected. For the RX1DAY index, The revised GLDAS data accurately captured extreme precipitation events in both time and space in 40 of the 45 years of records, an 63.4% improvement (from 0.215 to 0.849) over the accuracy before the correction (Figure 6).

4.2 | Data screening

It has been emphasized that the stability and independence of sequence data are the prerequisites for regional frequency analysis. Through data screening, we obtained the grids of regional frequency analysis in the annual maximum 1-day precipitation series 2,648 grids passed the tests. As Figure 7, the black grids were used for regional frequency analysis; the red grids that had not passed the test were excluded from the data set, and at-site frequency analysis was conducted during the subsequent processing.

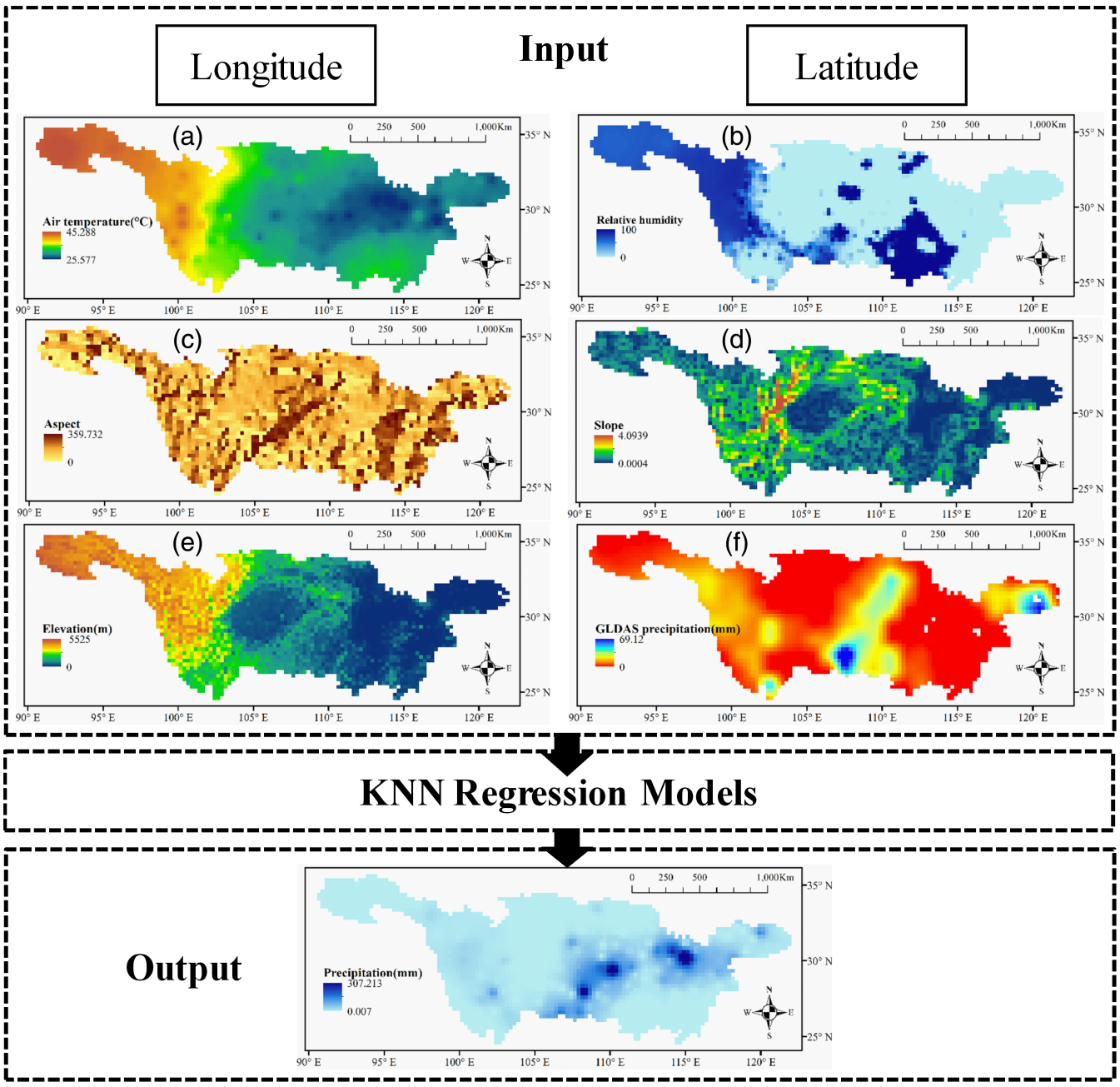

FIGURE 5 Correction of GLDAS precipitation by KNN regression with reference surfaces (a and b) from MicroMet model on July 22, 1998

TABLE 1 KNN models of three precipitation levels on July 22, 1998

Precipitation level	Stations	bestK	RMSE
Light rain	130	7	1.029
Moderate rain	25	10	2.852
Heavy rain	22	1	3.644

4.3 | Formation of homogeneous regions for RX1DAY series

The five standardized geographical features were selected for clustering, and the optimal clustering number of k-means clustering can be explored by the Davies-Bouldin index. Due to a large number of grids in the whole Yangtze River Basin, the spatial pattern of topography and annual precipitation in the cluster was considered, the upper limit of the cluster number range is set to 20. As

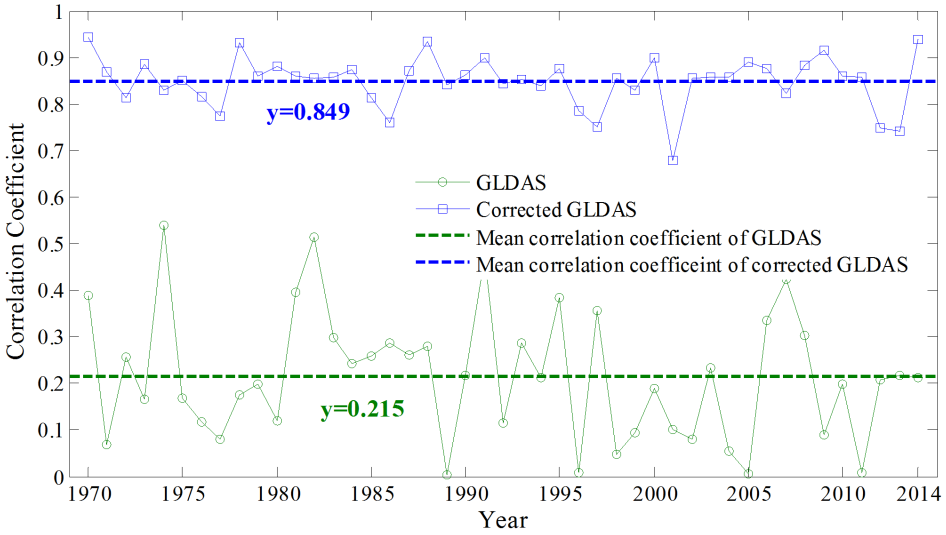

FIGURE 6 Correlation coefficient for RX1DAY between GLDAS (before and after correction) and ground observation

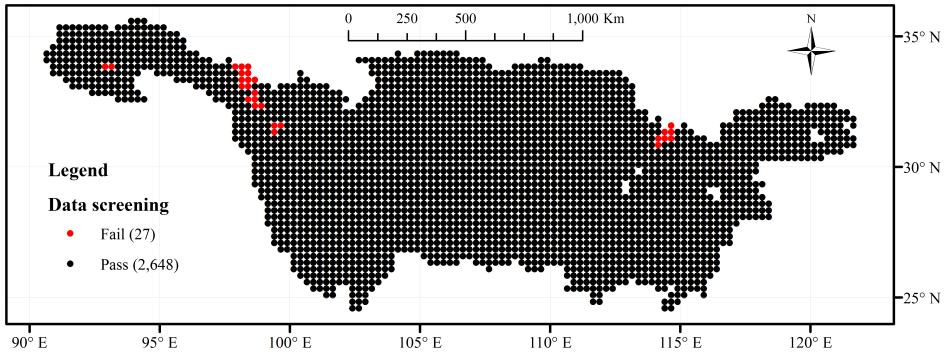

FIGURE 7 Result of data screening for the RX1DAY series. Empty (white) represent missing raw data

shown in Figure 8, the broken line reached a nadir when the number equalled seven; thus the optimal value of the initial clustering number was seven. Seven initial clusters were obtained by k-means clustering of all 2,675 grids in the Yangtze River Basin, these initial clusters were the prototype of the homogeneous precipitation regions.

The heterogeneity of seven initial clusters was evaluated by the discordancy measure and heterogeneity measure based on sample L-moments of each grid in the RX1DAY sequence. It can be seen from Table 2 that cluster 2 and cluster 3 were heterogeneous unless proper adjustments were made. Outliers were eliminated in small amounts (six grids) from Cluster 2 to form Region VI. After three iterative adjustments, Cluster 3 eliminated a total of 45 grids to obtain Region VII. All the H values after adjustments were less than two, which indicated that seven homogeneous regions of the annual maximum 1-day precipitation series in the Yangtze River Basin had been identified.

The result of the adjustment, the seven precipitation homogeneous regions for RX1DAY, is shown in Figure 9. These homogeneous regions in the Yangtze River Basin were generally spatially continuous; thus, the rationality and effectiveness of the partition could be accepted.

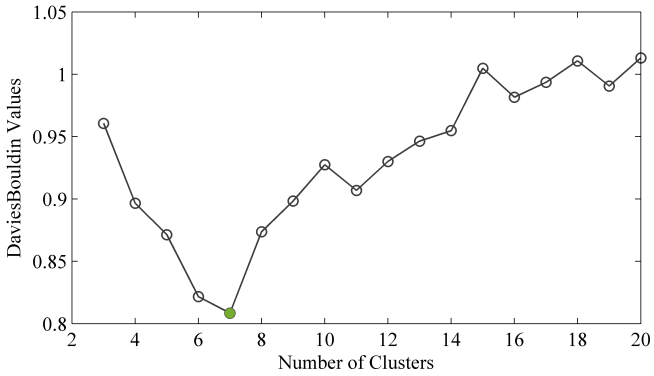

FIGURE 8 The optimal number of clusters from the Davies-Bouldin index

4.4 | Frequency analysis of RX1DAY series

A goodness of fit test was performed for the probability function of the regions (Table 3). GEV fitted well in most areas of the Yangtze River Basin, and GNO and GLO were also acceptable. For Regions III and VI, the three-parameter distribution could not be their acceptable

TABLE 2 Grid number and H value of initial clusters and final homogeneous regions for RX1DAY, bold part indicate clusters adjusted by automatic subjective adjustment

Cluster	Initial grids	*H* values	Region	Final grids	*H* value
1	424	1.992	V	424	1.992
2	**254**	**2.278**	**VI**	**248**	**1.63**
3	**390**	**5.212**	**VII**	**345**	**1.98**
4	288	0.899	I	288	0.899
5	436	1.197	II	436	1.197
6	461	−0.397	III	461	−0.397
7	395	−0.412	IV	395	−0.412

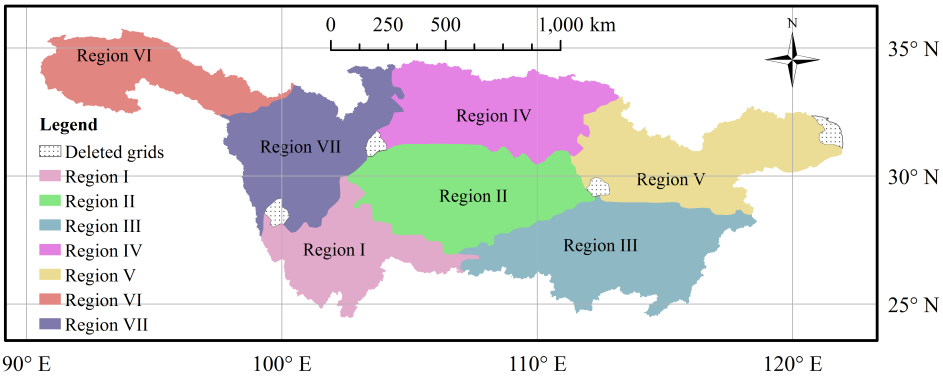

FIGURE 9 Seven homogeneous regions of RX1DAY series in the Yangtze River Basin

TABLE 3 Accept distributions and best distribution of seven homogeneous regions

Homogeneous region	Accept distribution	Best fitted distribution
I	GEV	GEV
II	GEV	GEV
III	WAK	WAK
IV	GEV GNO	GNO
V	GEV	GEV
VI	WAK	WAK
VII	GEV	GEV

distribution. Therefore, the five-parameter distribution Wakeby was chosen for further analysis.

The average correlation coefficient between sites in each homogeneous region is shown in Table 4. The average correlation coefficient between grids in the homogeneous regions ranged from 0.173 to 0.301, with a maximum of 0.301 in Region I. Therefore, the accuracy of the estimated regional growth curve should be evaluated on the basis of certain heterogeneity and inter-site dependence.

The estimated growth curve and estimation accuracy were obtained by using the optimal distribution of the precipitation homogeneous regions in the Yangtze River Basin and the annual maximum 1-day precipitation sequence of the grid in the region. Taking Region I, for example, L-CVs ranged from 0.119 to 0.225, L-kurtosis was 0.191, and the average H value of the simulation regions was 0.899. It can be ensured that the heterogeneity of the simulated region is equal to that of the actual region. The simulation region with 288 grids, 45-year precipitation record length, average sample L-moment ratio and inter-site correlation same as Region I was simulated 10,000 times by using the optimal fitting distribution GEV.

Similarly, Monte Carlo simulations were performed on the other homogeneous regions. The estimated regional growth curves and accuracy results of seven homogeneous regions are listed in Table 5. Considering the RMSE results and 99% error bound, it was concluded that the results of regional frequency analysis were very reliable for the quantile estimation of RX1DAY.

The regional growth curve and the 99% error upper and lower bounds of all seven homogenous regions are shown in Figure 10. It was also confirmed from the graph that the estimated quantiles are sufficiently reliable.

During the previous process, the RX1DAY series eliminated 27 grids because of data stationarity and independence and excluded 51 discordant grids for the formation of homogeneous regions. When analysing the spatial and temporal characteristics of extreme precipitation in the Yangtze River Basin, theoretically all grids should be

TABLE 4 Inter-site correlation coefficients of seven homogeneous regions

Region	I	II	III	IV	V	VI	VII
ρ	0.301	0.217	0.234	0.298	0.173	0.270	0.192

TABLE 5 Estimated regional growth curve and accuracy of seven homogeneous regions

	Region I				Region II			
Return period	$\hat{P}$	RMSE	Bound.0.01	Bound.0.99	$\hat{P}$	RMSE	Bound.0.01	Bound.0.99
2	0.945	0.006	0.932	0.958	0.931	0.005	0.918	0.942
5	1.212	0.012	1.186	1.235	1.209	0.009	1.187	1.226
10	1.394	0.023	1.344	1.444	1.412	0.019	1.39	1.453
20	1.573	0.036	1.494	1.655	1.618	0.032	1.551	1.692
50	1.811	0.057	1.686	1.943	1.909	0.053	1.802	2.036
100	1.995	0.076	1.828	2.171	2.145	0.073	2.001	2.323
Return period	Region III				Region IV			
	$\hat{P}$	RMSE	Bound.0.01	Bound.0.99	$\hat{P}$	RMSE	Bound.0.01	Bound.0.99
2	0.919	0.008	0.901	0.934	0.944	0.005	0.933	0.957
5	1.205	0.0103	1.181	1.224	1.230	0.010	1.208	1.251
10	1.435	0.023	1.384	1.487	1.420	0.020	1.375	1.464
20	1.679	0.039	1.599	1.772	1.602	0.031	1.532	1.673
50	2.029	0.065	1.897	2.187	1.838	0.048	1.730	1.948
100	2.313	0.091	2.131	2.535	2.017	0.062	1.877	2.159
Return period	Region V				Region VI			
	$\hat{P}$	RMSE	Bound.0.01	Bound.0.99	$\hat{P}$	RMSE	Bound.0.01	Bound.0.99
2	0.934	0.005	0.923	0.943	0.931	0.007	0.914	0.946
5	1.209	0.011	1.187	1.228	1.215	0.013	1.187	1.241
10	1.406	0.021	1.364	1.450	1.428	0.026	1.373	1.485
20	1.607	0.033	1.542	1.680	1.640	0.040	1.556	1.732
50	1.886	0.052	1.787	2.006	1.920	0.062	1.788	2.062
100	2.109	0.069	1.980	2.273	2.130	0.081	1.953	2.317
Return period	Region VII							
	$\hat{P}$	RMSE	Bound.0.01	Bound.0.99				
2	0.956	0.003	0.949	0.963				
5	1.194	0.009	1.175	1.210				
10	1.351	0.016	1.317	1.384				
20	1.501	0.024	1.451	1.554				
50	1.695	0.036	1.620	1.776				
100	1.840	0.046	1.745	1.945				

involved; thus, 78 grids were analysed vis at-site frequency analysis to obtain the estimated quantiles for each return period. Figure 11 shows the fitting distributions of the site (grid) 255 in the upper reaches of the Yangtze River and site (grid) 1971 in the middle and lower reaches of the Yangtze River, as well as the estimated quantile under a return period of 2–100 years.

4.5 | Spatial characteristics of extreme precipitation

When the estimated quantiles of all grids indices for RX1DAY in the Yangtze River Basin were obtained at the return period of 100 years, the geo-statistical interpolation method of ArcMap was used to describe the

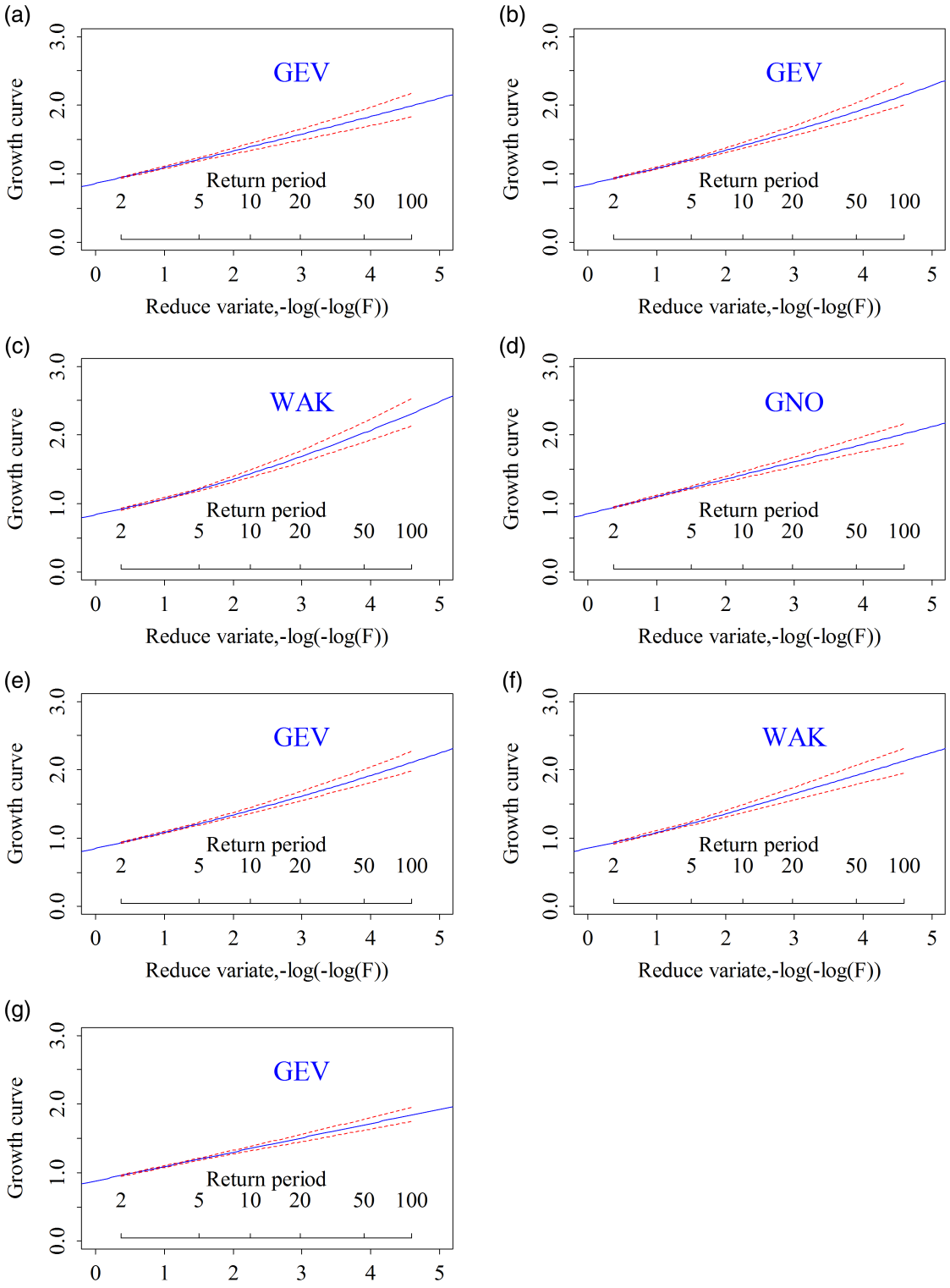

FIGURE 10 Regional growth curve (blue) and 99% error bounds (red) for seven sub-regions. Panels a–g are for regions I-VII of RX1DAY series, respectively

spatiotemporal characteristics of extreme precipitation in the study area. The ordinary Kriging interpolation was used to obtain an extreme precipitation spatial distribution map of the Yangtze River Basin (Figure 12).

Considering the typical monsoonal climate in the Yangtze River Basin (Zhang *et al.*, 2014), the topography is complex, and the west is high and the east is low in elevation. Reviewing the general situation in Figure 12, the

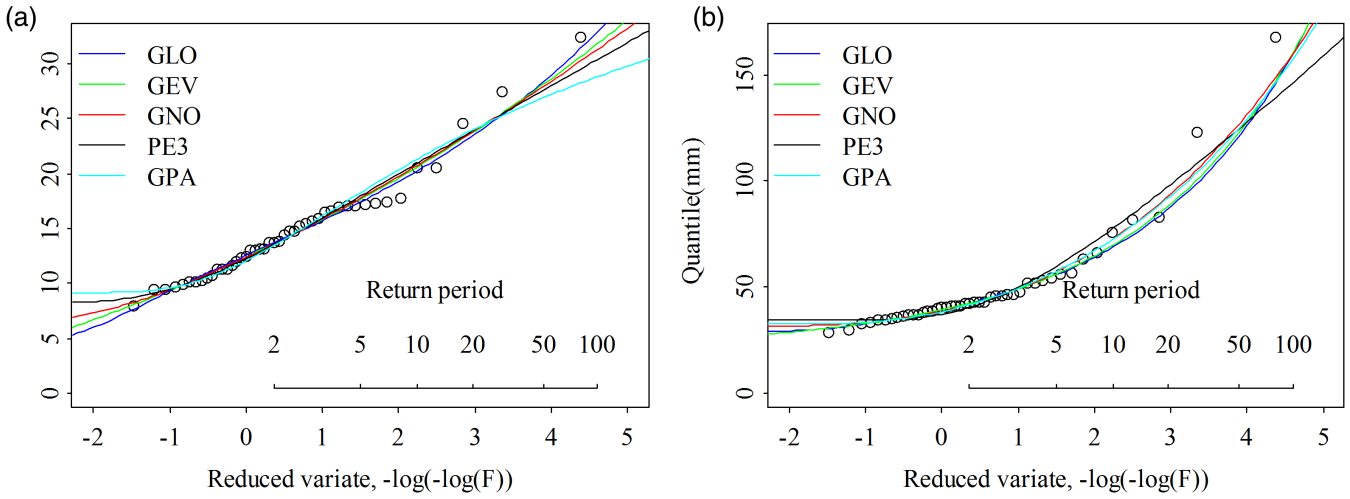

FIGURE 11 Extreme-value plot with fitted probability functions for RX1DAY at grid 255 (a) and grid 1971 (b)

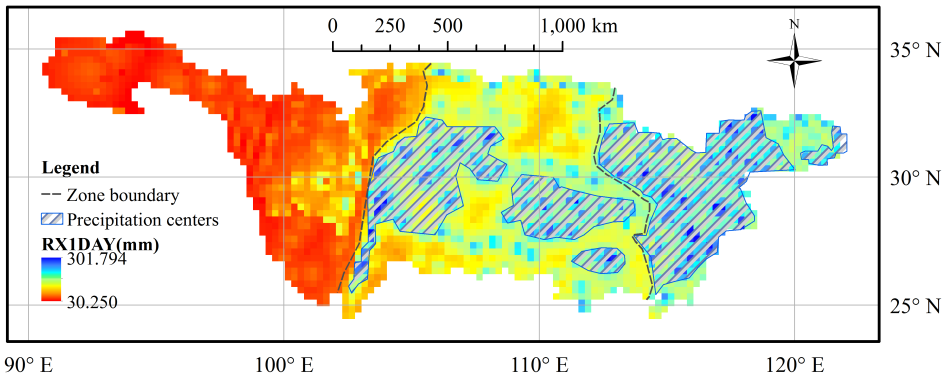

FIGURE 12 Spatial pattern of extreme precipitation with a return period of 100 years of RX1DAY

extreme precipitation distribution is uneven and can be approximately divided into three comparative zones: the "arid zone" in the upper reaches (the source area), the "wet zone" covering the most area of the basin, and the "special wet zone" (Fischer *et al*., 2012).

There were several precipitation centres (shaded areas in Figure 12) along the major rivers in the Yangtze River Basin. The biggest three were around the Sichuan Basin and the Dongting Lake Basin and area covered by a triangle (vertices are 120.161°E, 30.335°N; 112.611°E, 32.085°N, and 114.661°E, 25.585°N).

The western Sichuan Basin appeared as an obvious precipitation centre. Warm and humid airflows from the east and south are conducive to entering the basin and are blocked by the surrounding mountains; thus, extreme precipitation with of high intensity and long duration easily result in the closed high-temperature centre of the Sichuan Basin (Li *et al*., 2016).

The two precipitation centres located in the middle and lower reaches of the Yangtze River Basin include Dongting Lake Basin and the large triangle area, which covers the Poyang Lake Basin and South foot of Dabie Mountain. The middle and lower Yangtze River Basin is in a subtropical monsoonal climatic zone, and the East Asian summer monsoon is an important circulation system that affects the precipitation change (Pei *et al*., 2017). These precipitation trends and changes are also seen in the regional growth curve shown in Figure 10. The high altitude of the Dabie Mountain brings a lot of topographic rain. The increased precipitation can be associated with frequent climatic fluctuations or the stagnancy of the Meiyu rain belt; the greenhouse effect leads to increased temperature, as well as an increase in the climate variability of precipitation in this area. El Niño may result in an obvious plume of rain in the middle and lower reaches of the Yangtze River, which directly causes floods in the region (Wang and Yuan, 2018).

In the case of a 100-year return period, the designed maximum annual 1-day precipitation value ranged from 30.3 to 301.8 mm in the Yangtze River Basin, with great differences in the area. Combined with the spatial distribution, we should pay attention to the rainfall intensity of the rainstorm area near the northeast of Jiangxi province and the rainstorm area of Sichuan Basin, as well as

the precipitation increment of the middle and lower reaches affected by the Meiyu and flood season (Su *et al*., 2008).

Compared to the previous results, which were based on six approximate zones (Chen *et al*., 2014b), the precipitation extremes characteristics obtained in this paper show general consistency but a higher spatial resolution. The results proved to some extent the reliability of the automatic subjective adjustment method. The boundary of the precipitation centre appeared much smoother and could more accurately represent the range of each precipitation centre, probably thanks to the correction of the GLDAS daily precipitation dataset (Qi *et al*., 2018).

5 | CONCLUSIONS

Regional frequency analysis based on corrected satellite-based grid precipitation data was conducted to obtain the spatiotemporal characteristics of the Yangtze River Basin in the context of the water resource and flood risk management at a basin scale. According to the results of this study, we can draw several conclusions as follows:

1. The corrected GLDAS daily precipitation data had greatly improved its ability to capture extreme precipitation events in the Yangtze River Basin, as the data average accuracy increased from 0.215 before correction to 0.849 after correction. It is feasible and reliable to use corrected GLDAS data to replace the station data for the regional frequency analysis of extreme precipitation.
2. The effectivity of the automatic subjective adjustment method proposed in this paper was proved by the seven detailed homogeneous regions within the Yangtze River Basin. This method provides a basis for efficient calculation of regional frequency analysis with satellite-based grid precipitation products.
3. The regional growth curves and quantiles of the Yangtze River Basin were derived for the return period for 2–100 years. The RMSEs and the 99% error bounds derived via Monte Carlo simulation showed that the regional growth curves and quantiles were reliable sufficient according to the inter-site correlation of homogeneous regions.
4. Spatial patterns of extreme daily precipitation series with a return period of 100 years indicated that the precipitation amount increases gradually from the upper to the lower Yangtze River Basin, from the "arid zone" to the "wet zone" and then to the "special wet zone", and the 100-year return level of maximum 1-day precipitation varied from 30.3 to 301.8 mm. There were three main precipitation centres, which concluded the Sichuan Basin, Dongting Lake Basin, and great triangle area covering the Poyang Lake Basin and the south foot of Dabie Mountain, with the high risk of flood disaster caused by extreme precipitation.

The geographic factors used in the clustering procedure may have inaccurately judged precipitation. In the future, more appropriate indicators need be further studied.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (grant numbers 41771422, 41971351, and 41890822).

ORCID

Yi Zeng https://orcid.org/0000-0002-8007-3415
Lei Xu https://orcid.org/0000-0002-6454-2963

REFERENCES

Alvarez, M., Puertas, J. and Pena, E. (2016) Regional frequency analysis of extremes precipitations in northern of Mozambique. *Ingenieria del agua*, 20, 29–42.

Anli, A.S., Apaydin, H. and Ozturk, F. (2009) Regional frequency analysis of the annual maximum precipitation observed in Trabzon Province. *Journal of Agricultural Sciences-Tarim Bilimleri Dergisi*, 15, 240–248.

Bai, P. and Liu, X.M. (2018) Intercomparison and evaluation of three global high-resolution evapotranspiration products across China. *Journal of Hydrology*, 566, 743–755.

Basheer, M. and Elagib, N.A. (2019) Performance of satellite-based and GPCC 7.0 rainfall products in an extremely data-scarce country in the Nile Basin. *Atmospheric Research*, 215, 128–140.

Bregt, A.K., Mcbratney, A.B. and Wopereis, M.C.S. (1991) Construction of isolinear maps of soil attributes with empirical confidence-limits. *Soil Science Society of America Journal*, 55, 14–19.

Chen, J., Wu, X., Finlayson, B.L., Webber, M., Wei, T., Li, M. and Chen, Z. (2014a) Variability and trend in the hydrology of the Yangtze River, China: annual precipitation and runoff. *Journal of Hydrology*, 513, 403–412.

Chen, Y.D., Zhang, Q., Xiao, M.Z., Singh, V.P., Leung, Y. and Jiang, L.G. (2014b) Precipitation extremes in the Yangtze River basin, China: regional frequency and spatial-temporal patterns. *Theoretical and Applied Climatology*, 116, 447–461.

ETCCDI/CRD, 2013. Climate Change Indices [Cited April 8, 2020]. http://etccdi.pacificclimate.org/list_27_indices.shtml.

Fischer, T., Su, B.D., Luo, Y. and Scholten, T. (2012) Probability distribution of precipitation extremes for weather-index based insurance in the Zhujiang River basin, South China. *Journal of Hydrometeorology*, 13, 1023–1037.

Gao, T. and Xie, L. (2016) Spatiotemporal changes in precipitat ion extremes over Yangtze River basin, China, considering the rainfall shift in the late 1970s. *Global and Planetary Change*, 147, 106–124.

Gao, T., Xie, L. and Liu, B. (2016) Association of extreme precipitati on over the Yangtze River basin with global air–sea heat fluxes and moisture transport. *International Journal of Climatology*, 36, 3020–3038.

Gao, F., Zhang, Y.H., Chen, Q.H., Wang, P., Yang, H.R., Yao, Y.J. and Cai, W.Y. (2018) Comparison of two long-term and high-resolution satellite precipitation datasets in Xinjiang, China. *Atmospheric Research*, 212, 150–157.

Gemmer, M., Jiang, T., Su, B.D. and Kundzewicz, Z.W. (2008) Seasonal precipitation changes in the wet season and their influence on flood/drought hazards in the Yangtze River basin, China. *Quaternary International*, 186, 12–21.

Gnanadesikan, R., Kettenring, J.R. and Maloor, S. (2007) Better alternatives to current methods of scaling and weighting data for cluster analysis. *Journal of Statistical Planning and Inference*, 137(11), 3483–3496.

Goyal, M.K. and Gupta, V. (2014) Identification of homogeneous rainfall regimes in northeast region of India using fuzzy cluster analysis. *Water Resources Management*, 28(13), 4491–4511.

Guo, J., Guo, S., Li, Y., Chen, H. and Li, T. (2013) Spatial and temporal variation of extreme precipitation indices in the Yangtze River basin, China. *Stochastic Environmental Research & Risk Assessment*, 27(2), 459–475.

Hirsch, R.M., Slack, J.R. and Smith, R.A. (1982) Techniques of trend analysis for monthly water-quality data. *Water Resources Research*, 18(1), 107–121.

Hlavcova, K., Lapin, M., Valent, P., Szolgay, J., Kohnova, S. and Roncak, P. (2015) Estimation of the impact of climate change-induced extreme precipitation events on floods. *Contributions to Geophysics and Geodesy*, 45(3), 173–192.

Hosking, J.R.M. and Wallis, J.R. (1993) Some statistics useful in regional frequency analysis. *Water Resources Research*, 29, 271–282.

Hosking, J.R.M. and Wallis, J.R. (1997) *Regional Frequency Analysis*. Cambridge: United Kingdom.

Hosking, J.R.M., Wallis, J.R. and Wood, E.F. (1985) Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27(3), 251–261.

Jiang, T., Kundzewicz, Z.W. and Su, B.D. (2010) Changes in monthly precipitation and flood hazard in the Yangtze River Basin, China. *International Journal of Climatology*, 28, 1471–1481.

Koster, R.D., Suarez, M.J., Ducharne, A., Stieglitz, M. and Kumar, P. (2000) A catchment-based approach to modeling land surface processes in a general circulation model 1. Model structure. *Journal of Geophysical Research-Atmospheres*, 105, 24809–24822.

Kunkel, K.E. (1989) Simple procedures for extrapolation of humidity variables in the mountainous Western United States. *Journal of Climate*, 2(7), 656–669.

Li, D.S., Sun, J.H., Fu, S.M., Wei, J., Wang, S.G. and Tian, F.Y. (2016) Spatiotemporal characteristics of hourly precipitation over central eastern China during the warm season of 1982-2012. *International Journal of Climatology*, 36(8), 3148–3160.

Li, B., Beaudoing, H., and Rodell, M. (2018) GLDAS Catchment Land Surface Model L4 daily 0.25 x 0.25 degree V2.0. https://doi.org/10.5067/LYHA9088MFWQ.

Li, J., Wang, Z., Wu, X., Xu, C., Guo, S. and Chen, X. (2020) Toward monitoring short-term droughts using a novel daily scale, standardized antecedent precipitation evapotranspiration index. *Journal of Hydrometeorology*, 21, 891–908.

Liang, Y.Y., Liu, S.G., Guo, Y.P. and Hua, H. (2017) L-moment-based regional frequency analysis of annual extreme precipitation and its uncertainty analysis. *Water Resources Management*, 31(12), 3899–3919.

Liston, G.E. and Elder, K. (2006) A meteorological distribution system for high-resolution terrestrial modelling (MicroMet). *Journal of Hydrometeorology*, 7(2), 217–234.

Lü, M., Wu, S.J., Chen, J., Chen, C., Wen, Z. and Huang, Y. (2018) Change s in extreme precipitation in the Yangtze River basin and its association with global mean temperature and ENSO. *International Journal of Climatology*, 38, 1989–2005.

Maulik, U. and Bandyopadhyay, S. (2002) Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12), 1650–1654.

Mehdizadeh, S. (2020) Using AR, MA, and ARMA time series models to improve the performance of MARS and KNN approaches in monthly precipitation Modeling under limited climatic data. *Water Resources Management*, 34(1), 263–282. https://doi.org/10.1007/s11269-019-02442-1.

Muhammad, E., Muhammad, W., Ahmad, I., Muhammad Khan, N. and Chen, S. (2020) Satellite precipitation product: applicability and accuracy evaluation in diverse region. *Science China Technological Sciences.*, 63, 819–828. https://doi.org/10.1007/s11431-019-1457-3.

Pei, F.S., Wu, C.J., Qu, A.X., Xia, Y., Wang, K. and Zhou, Y. (2017) Changes in extreme precipitation: a case study in the middle and lower reaches of the Yangtze River in China. *Water*, 9 (12), 943.

Qi, W., Liu, J.G. and Chen, D.L. (2018) Evaluations and improvements of GLDAS2.0 and GLDAS2.1 forcing Data's applicability for basin scale hydrological simulations in the Tibetan plateau. *Journal of Geophysical Research-Atmospheres*, 123(23), 13128–13148.

Risser, M.D., Paciorek, C.J., Wehner, M.F., O'Brien, T.A. and Collins, W.D. (2019) A probabilistic gridded product for daily precipitation extremes over the United States. *Climate Dynamics*, 53(5–6), 2517–2538.

Seyyedi, H., Anagnostou, E.N., Beighley, E. and McCollum, J. (2015) Hydrologic evaluation of satellite and reanalysis precipitation datasets over a mid-latitude basin. *Atmospheric Research*, 164, 37–48.

Silvestre, M.R., Flores, E.F. and Sant'Anna Neto, J.L. (2016) Geoestatistics applied to spatial distribution of the precipitation. *Revista Formacao Online*, 3(23), 317–338.

Su, B.D., Marco, G., Jiang, T. and Ren, G.Y. (2008) Probability distribution of precipitation extremes over the Yangtze River basin. *Advances in Climate Change Research*, 4, 27–31.

Sun, H.M., Wang, G.J., Li, X.C., Chen, J., Su, B.D. and Jiang, T. (2017) Regional frequency analysis of observed sub-daily rainfall maxima over eastern China. *Advances in Climate Change Research*, 34, 209–225.

Tang, G.Q., Clark, M.P., Papalexiou, S.M., Ma, Z.Q. and 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 Sensing of Environment*, 240, 111697. https://doi.org/10.1016/j.rse.2020.111697.

Usman, M., Nichol, J.E., Ibrahim, A.T. and Buba, L.F. (2018) A spatio-temporal analysis of trends in rainfall from long term satellite rainfall products in the Sudano Sahelian zone

of Nigeria. *Agriculture and Forest Meteorology*, 260, 273–286.

Wang, S.S. and Yuan, X. (2018) Extending seasonal predictability of Yangtze River summer floods. *Hydrology and Earth System Sciences*, 22(8), 4201–4211. https://doi.org/10.5194/hess-22-4201-2018.

Wang, W.G., Xing, W.Q., Yang, T., Shao, Q.X., Peng, S.Z., Yu, Z.B. and Yong, B. (2013) Characterizing the changing behaviours of precipitation concentration in the Yangtze River basin, China. *Hydrological Processes*, 27, 3375–3393. https://doi.org/10.1002/hyp.9430.

Wang, Y.D., Nan, Z.T., Chen, H. and Wu, X.B. (2016) Correction of daily precipitation data of ITPCAS dataset over the QingHai-Tibetan plateau with KNN model. *IEEE International Geoscience and Remote Sensing Symposium*, 593–596. https://doi.org/10.1109/IGARSS.2016.7729148.

Wright, P.B., Wallace, J.M., Mitchell, T.P. and Deser, C. (1988) Correlation structure of the El Nino/southern oscillation phenomenon. *Journal of Climate*, 1, 609–625.

Wu, X., Guo, S., Yin, J., Yang, G., Zhong, Y. and Liu, D. (2018a) On the event-based extreme precipitation across China: time distribution patterns, trends, and return levels. *Journal of Hydrology*, 562, 305–317.

Wu, X., Guo, S., Liu, D., Hong, X., Liu, Z., Liu, P. and Chen, H. (2018b) Characterization of rainstorm modes along the upper mainstream of Yangtze River during 2003-2016. *International Journal of Climatology*, 38, 1976–1988.

Xia, J., Huang, G.H., Chen, Z. and Rong, X. (2001) An integrated planning framework for managing flood-endangered regions in the Yangtze River basin. *Water International*, 26, 153–161.

Yang, T., Shao, Q.X., Hao, Z.C., Chen, X., Zhang, Z.X., Xu, C.Y. and Sun, L. (2010) Regional frequency analysis and spatio-temporal pattern characterization of rainfall extremes in the Pearl River Basin, China. *Journal of Hydrology*, 380, 386–405.

Yang, F., Lu, H., Yang, K., He, J., Wang, W., Wright, J.S., Li, C., Han, M. and Li, Y. (2017) Evaluation of multiple forcing data sets for precipitation and shortwave radiation over major land areas of China. *Hydrology and Earth System Sciences*, 21(11), 5805–5821.

Yin, Y., Chen, H., Xu, C.Y., Xu, W., Chen, C. and Sun, S. (2016) Spatio-temporal characteristics of the extreme precipitation by L-moment-based index-flood method in the Yangtze River Delta region, China. *Theoretical and Applied Climatology*, 124, 1005–1022.

Zhang, Q., Peng, J.T., Xu, C.Y. and Singh, V.P. (2014) Spatiotemporal variations of precipitation regimes across Yangtze River basin, China. *Theoretical and Applied Climatology*, 115, 703–712.

Zhang, Y., Hong, Y., Wang, X.G., Gourley, J.J., Xue, X.W., Saharia, M., Ni, G.H., Wang, G.L., Huang, Y., Chen, S. and Tang, G.Q. (2015) Hydrometeorological analysis and remote sensing of extremes: was the July 2012 Beijing flood event detectable and predictable by global satellite observing and global weather Modeling systems? *Journal of Hydrometeorology*, 16, 381–395.

Zhou, X.Y., Zhang, Y.Q., Yang, Y.G., Yang, Y.M. and Han, S.M. (2013) Evaluation of anomalies in GLDAS-1996 dataset. *Water Science and Technology*, 67, 1718–1727.

How to cite this article: Chen Z, Zeng Y, Shen G, Xiao C, Xu L, Chen N. Spatiotemporal characteristics and estimates of extreme precipitation in the Yangtze River Basin using GLDAS data. *Int J Climatol*. 2021;41 (Suppl. 1): E1812–E1830. https://doi.org/10.1002/joc.6813

APPENDIX A.

A. Algorithm for automatic subjective adjustment method

Input: Initial regions of precipitation series (RX1DAY) after K-means clustering.

Output: Homogeneous regions and deleted grids of each series.

Steps:

1. Calculate the heterogeneity of each initial cluster, record the homogeneous regions *Homs* and heterogeneous regions *Hts*, respectively.
2. The variable grid g is derived from the descending order of D values of grids in the heterogeneous region r (and probably are not discordant).
3. The heterogeneity of region r is H. Does the D value of grid g exceed the threshold ($D_g > 3 \parallel D_g > (N-1)/3$)?
   - 3.1. YES: Delete the grid g directly, calculate the heterogeneity value $H^{'}$, and compare with H before deletion.
     - 3.1.1. $H^{'} < 2$: The delete operation has completed and the adjustment of region r has ended, record the new homogenous region r and deleted grids gs;
     - 3.1.2. $H^{'} > 2$ & $H^{'} < H$: Repeat this step;
     - 3.1.3. $H^{'} > 2$ & $H^{'} > H$ (The H value does not decrease or increase in the process of deleting grid): Cancel all deletion operations, output the prompt "This cluster is heterogeneous!", and the adjustment of region r has ended;
   - 3.2. NO: Go to step 5.
4. Is grid g at the boundary of region r to which it belongs?
   - 4.1. Yes: Start from step 6;
   - 4.2. No: Delete grid g;
5. Region have too few grids (< 50) or region are staggered with other region(s), undo all deletion operations, merge, and repartition the original cluster. Check heterogeneity, $H^{'} < 2$?

5.1. $H^{'} < 2$: Merge and repartition operations are completed, and the adjustment of region has ended;
5.2. $H^{'} > 2$ & $H^{'} < H$: Repeat step 4;
5.3. $H^{'} > 2$ & $H^{'} > H$: Cancel the merge operation, and output the "All of these zones are heterogeneous regions!" prompt, and the adjustment of region r has ended;

6. Move grid g. Calculate the H values of both move-out and move-in zones.
6.1. $H_{out} < 2$ & $H_{in} < 2$: The move operation has completed, and the adjustment of region r has ended;
6.2. $H_{out} < 2 \parallel H_{in} < 2$: Repeat step 4;
6.3. $H_{out} > 2$ & $H_{in} > 2$: Cancel the move operation;

7. Repeat 1–6 until each region is identified as a homogeneous/heterogeneous region, record the final homogeneous regions *Homs* and final deleted grids *Gs* (which contain the heterogeneous regions), the subjective adjustment ends.