[1]Xin GUO
[2]Zhijie Yin
[3*]Cheng Gao
[4]Boyao Zhang
[5]Manqiu Hao
[6]Xiaomin Ji

Frequency Analysis of Rainfall with Different Methods based on the Missing Data Processed by Random Forest Algorithm

Abstract: - Global warming is causing dramatic climate change, leading to rainfall that triggers a more severe risk of flooding. The conventional moment method and linear moment method's basic theories for the design of flood frequency analyses were introduced and compared for frequency analysis of rainfall. The rainfall data from 149 long-series representative rainfall stations in Jiangsu Province were processed using the random forest (RF) algorithm to address the missing data encountered during the actual rainfall monitoring process. The frequency was calculated using conventional and linear moment methods, and the differences and advantages of the two methods were analyzed by comparing the calculation results of different sites under each method. The results show that under low design frequencies, both the conventional moment and linear moment methods exhibit minimal errors, rendering them suitable for calculating design rainfall. The linear moment method outperforms the conventional moment method in terms of the unbiasedness of the estimation process and for very large values, and that the parameters estimated by the linear moment method are more accurate. In practical hydrological frequency calculations, different computation methods can be chosen according to specific needs to enhance calculation accuracy.

Keywords: Hydrological Frequency Calculation, Linear Moment Method, Conventional Moment Method, P-III Distribution Frequency Curve, Random Forest Algorithm.

I. INTRODUCTION

Floods have always been one of the most serious and malevolent natural disasters, and countries around the world are threatened by floods to varying degrees, attracting widespread concern and attention from the government and academia [1-2]. Global warming has led to significant shifts in both average and extreme weather conditions across numerous areas, modifying the spatial and temporal patterns of water resource distribution and the water cycle. These changes have profound effects on agricultural output, worldwide biodiversity, and the overall human existence [3-4]. China is a flood-prone country, and flood prevention is a critical issue concerning safety of people's lives, property, and social stability. Over the last two decades, China's cities have experienced frequent flooding, and the phenomena of blockage, flooding, and heavy rainfall have become increasingly serious in most cities. Most floods in China are caused by torrential rain. In China, one important method for calculating design floods is to project design floods based on torrential rainfall data. Leveraging machine learning techniques, hydrological prediction and simulation have become essential tools for accurately forecasting and managing water resources, mitigating water-related disasters, and optimizing water utilization strategies by efficiently processing and analyzing vast amounts of meteorological and hydrological data, thus enhancing the accuracy of predictions and the efficiency of simulations. By establishing machine learning models, it is possible to predict hydrological variables such as rainfall, flood flow, and reservoir storage. Common machine learning models include random forest, neural networks, and support vector machines. These models can utilize historical data and meteorological information to predict future hydrological conditions, providing decision support for water resource management and flood disaster prevention. However, they also have the drawback of being prone to overfitting [5-6]. To address the overfitting issue, Breiman improved the regression tree model based on the bagging method and proposed the random forest model [7]. The Random Forest algorithm functions by building numerous decision trees in the training phase and then compiling their outcomes for predictions. This approach not only improves the accuracy of the model but also effectively reduces the problem of overfitting. By training multiple decision trees and combining their predictive outcomes, Random Forest is capable of capturing complex relationships within the data,

[1] College of Hydrology and Water Resources, Hohai University, Jiangsu Nanjing 210098, China
[2] Information Center, Ministry of Water Resources, Beijing 100053, China
[3] College of Hydrology and Water Resources, Hohai University, Jiangsu Nanjing 210098, China
[4] College of Hydrology and Water Resources, Hohai University, Jiangsu Nanjing 210098, China
[5] College of Hydrology and Water Resources, Hohai University, Jiangsu Nanjing 210098, China
[6] Jiangsu Province Hydrology and Water Resources Investigation Bureau, Jiangsu Suzhou 215129, China
*Corresponding author: Cheng Gao

while maintaining a high level of generalization to new data. It is widely applied in fields such as hydrology and geography [8-9].

However, in practice applications, frequency analysis calculation remains the most common and effective method of determining the design flood. It is a critical basis for the design of hydraulic facilities and flood control management measures, especially within the framework of ongoing climate change, and is essential for identifying extreme values with a specified likelihood of occurrence [10-11]. Hydrologic frequency analysis began about 1880~1890, Herschel and Rafter in the United States first applied the frequency curve (then called the duration curve); in 1896, Horton applied the frequency analysis method to the runoff study; in 1913~1914, Fuller and Hazen published papers describing the application of the frequency method; in 1921, Hazen proposed the use of logarithmic lattice probability paper and began fitting lines on it, which was the first lognormal distribution application[12]. Hydrological frequency calculations are the theoretical basis of China's flood control design standards and are an important link in the design of water conservancy projects and water resource management. The frequency calculation method is mainly used to analyze the statistical change characteristics of a flood peak volume (or rainfall extreme value), explore the quantitative relationship between the frequency and flood peak volume, and deduce the design value of the flood peak volume (or rainfall extreme value) for a certain design return period. The primary concern in flood-frequency calculations are sampling methods, parameter estimation and line selection. The overall frequency distribution curve's line shape is unknown, and a line shape is usually chosen to provide a good fit to most hydrologic series. A distribution curve is generally defined by a small number of parameters.

Before calculating the flood frequency, we must first choose a statistical distribution model for the hydrological series. According to the Chinese flood data site, the Pearson type III (P-III) distribution curve has been considered suitable for most flood series in China since the 1960s. Therefore, Flood Calculation Specification for Water Conservancy and Hydropower Engineering Design (SL44-2006) states that "P-III shall be used for the line type of frequency curve, and other line types may be used after demonstration for special cases" (SL44-2006). In addition, because flood characteristics differ by location, distribution curves, including the generalized extreme value distribution (GEV), lognormal distribution (CNO), and generalized Pareto distribution (GPA), are recommended for flood frequency analysis at home and abroad [13–15]. The second method estimates the parameters in the distribution model, and the current methods of parameter estimation include the method of moments, method of weight functions, method of probability weight moments, linear moments, visual estimation of the appropriate line, and computer optimization of the appropriate line [16–18]. Hosking (1990) proposed a linear moment method based on probabilistic weight moments, and its good unbiasedness and robustness as a new parameter estimation method have piqued the interest of academic and engineering communities[19,20]. Since 1990, the Office of Hydrology (OHD) under the National Oceanic and Atmospheric Administration (NOAA) of the United States has conducted research on the application of the zonal linear moment method in flood control design standards. In 2006, the Office proposed a complete system for analyzing rainstorm frequency using the linear moment method combined with a regional comprehensive analysis method, which has been promoted nationwide in the United States. The calculation of rainstorm frequency in the United States was conducted and incorporated into the national standards for flood management. Several domestic and international scholars have compared and analyzed different aspects of the conventional moment method and the linear moment method. Sankarasubramanian and Srinivasan [21] verified the superiority of the linear moment method over the conventional moment method by using actual data from several stations in India's central region. Fill and Stedinger [22] conducted a comparative study of linear moments and constant rules and concluded that the quantile test estimated using the linear moment parameter is more effective than the conventional moment method. Hussain et al. [23] analyzed seven sites in the Pakistan's Punjab area using a regional synthesis method based on the linear method of moments, and chose the most robust distribution. Liang et al. [24] used the Taihu Lake Basin in China as an example to demonstrate the theoretical superiority of the linear moment method over the conventional moment method in a preliminary comparison. Anghel and Ilinca [25] presented improved approximations for the estimation of probability distributions of hydrological extremes using the conventional moment method and the linear moment method and, in some cases, new approximations that provided a new paradigm for updating the normative standards in the field of hydraulics in Romania.

In this paper, the estimation results of the conventional moment method and the linear moment method are analyzed and evaluated by utilizing the rainfall data of Jiangsu Province that has been processed by the random forest algorithm, along with the most representative P-III curve in China.

II. Algorithms and Theoretical Comparison of Conventional Moment Method and Linear Moment Method

Random Forest is a technique that utilizes several untrimmed classification and regression decision trees for both prediction and classification tasks. Its core concept is that each tree relies on a sample randomly drawn from the original dataset, and features are selected randomly during the growth process of the tree. The method of moments (MOM), a classical approach, is utilized for parameter estimation. For most distribution functions, moments of all orders of origin and central moments exist. Moreover, a correlation is present between the moments and the parameters of the distribution function, allowing for the representation of parameters through moments. The method of moments estimates the statistical parameters of a frequency curve by replacing (or estimating) the overall moments with the sample moments and using the equation of the relationship between the moments and parameters.

A. Random Forest Algorithm

Random forests employ the Bootstrap resampling technique to randomly select samples from the original dataset and construct decision trees for each sample set, resulting in a series of classification models {h1(X), h2(X),…hn(X)}; n different results are obtained from the predictions made by the established series of classification models, and the final prediction result is obtained by voting or taking the average [26].

The specific process is as follows:

Step 1 uses bootstrapping to randomly generate K distinct sample datasets from the original dataset, which act as the sub-training set for each decision tree. The size of each sample matches that of the original dataset, with the data not selected in each sampling forming the out-of-bag data.

Step 2 constructs a classification regression tree for each sample dataset, generating K decision trees. During the generation process, for each node within the decision tree, a subset of variables is obtained through random sampling from the original data variable set. The best variable for node splitting and branching is chosen from the subset by aiming to minimize the Gini index, according to the selection criterion.

Step 3 every classification regression tree iteratively bifurcates from the top downwards until it meets the predefined minimum size for leaf nodes, termed nodesize, halting further expansion of the decision tree. These decision trees are then aggregated to create a random forest.

Step 4 feeds the test data into model, employing the K decision trees for individual predictions, and calculates the mean of the outcomes from each decision tree to determine the regression value, that is, the forecasted value.

This method boosts the model's capacity to generalize, lowers the likelihood of overfitting, and amalgamates the individual decision trees' classification and regression outcomes, counterbalancing certain random inaccuracies. [27]. The flowchart of the calculation is depicted in Figure 1.

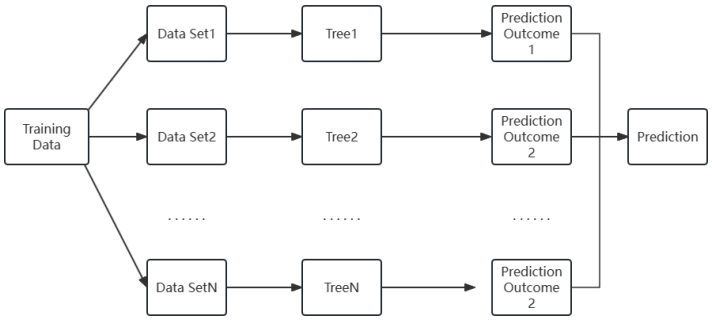


Figure 1: Flowchart of Random Forest Calculation

B. Conventional Moment Method

The conventional moment method, a classical approach for parameter estimation broadly employed, substitutes overall moments with sample moments.For the vast majority of distribution functions, there are origin moments and center moments, and the moments and distribution parameters tend to have a certain relationship; thus, it is common to use moments to represent each parameter to establish the corresponding relationship to estimate the parameters of the frequency curve. For the probability density function f(x) of a distribution, we define its moments of origin of order r as follows:

$$\mu_r' = \int_{-\infty}^{\infty} x^r f(x) dx \quad (1)$$

Its r-order center moments are as follows:

$$\mu_r = \int_{-\infty}^{\infty}(x-\mu_1')^r f(x)dx \tag{2}$$

The unbiased estimates of the statistical parameters obtained from the method of moments are:

$$\bar{x} = \mu_1' = mean \tag{3}$$

$$C_v = \frac{\sqrt{\mu_2}}{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{n}(K_i-1)^2}{n-1}} \tag{4}$$

$$C_s = \frac{n\sum_{i=1}^{n}(K_i-1)^3}{(n-1)(n-2){C_v}^3} \tag{5}$$

This technique is easy to compute and ranks among the most frequently employed methods for conducting frequency analysis calculations; however, it has significant drawbacks. According to previous research results, the moment method produces statistical parameters and estimated frequencies that are notably lower. Therefore, the obtained Cv and Cs are clearly insufficient, and the accuracy of the obtained design values must be improved. Typically, the values obtained directly from this method is used as references.

C. Linear Moment Method

Hosking characterized Linear moments as the expected values derived from specific linear combinations of order statistics, alongside linear coefficients of divergence and additional higher-order linear moment ratios, in contrast to the traditional method of moments. As a new parameter estimation method, the linear moment method has good unbiasedness and robustness to very large values as a new parameter estimation method. It was calculated as follows:

Assuming that the variable X follows a certain distribution function, defining the r-order linear moment variable as:

$$\lambda_r \equiv r^{-1}\sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}EX_{r-k:r}, r = 1,2,\cdots \tag{6}$$

By definition, the first four linear moments of this random variable X are:

$$\begin{gathered}\lambda_1 = EX\\ \lambda_2 = \frac{1}{2}E(X_{2:2}-X_{1:2})\\ \lambda_3 = \frac{1}{3}E(X_{3:3}-2X_{2:3}+X_{1:3})\\ \lambda_4 = \frac{1}{4}E(X_{4:4}-3X_{3:4}+3X_{2:4}-X_{1:4})\end{gathered} \tag{7}$$

The first four orders of sample linear moments for discrete samples are:

$$\begin{gathered}l_1 = n^{-1}\sum_{i=1}^{n}x_i\\ l_2 = \frac{1}{2}\binom{n}{2}^{-1}\sum_{i=j+1}^{n}\sum_{j=1}^{n-1}(x_{i:n}-x_{j:n})\\ l_3 = \frac{1}{3}\binom{n}{3}^{-1}\sum_{i=j+1}^{n}\sum_{j=k+1}^{n-1}\sum_{k=1}^{n-2}(x_{i:n}-2x_{j:n}+x_{k:n})\\ l_4 = \frac{1}{4}\binom{n}{4}^{-1}\sum_{i=j+1}^{n}\sum_{j=k+1}^{n-1}\sum_{k=l+1}^{n-2}\sum_{l=1}^{n-3}(x_{i:n}-3x_{j:n}+3x_{k:n}-x_{l:n})\end{gathered} \tag{8}$$

The general form of the r-order sample linear moments for discrete samples is:

$$l_r = r^{-1}\binom{n}{r}^{-1}\sum_{n\geq i_r=i_{r-1}+1}^{n}\cdots\sum_{i_2=i_1+1}^{n-r+2}\sum_{i_1=1}^{n-r+1}*\sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}x_{i_{r-k:n}}, r = 1,2,\cdots,n \tag{9}$$

For sample linear moments, the expected value of the order statistic is defined as:

$$EM_{r:n} = \frac{n!}{(r-1)!(n-r)!}\int_0^1 x[F(x)]^{r-1}[1-F(x)]^{n-r}dF(x) \tag{10}$$

The probability weight moments are defined as:

$$M_{i,j,k} = \int_0^1 x^i F^j(1-F)^k dF \tag{11}$$

Where i, j and k are the order of moments, all positive integers. In order to avoid large sampling errors from higher orders, i=1, j=0 or k=0 is usually taken to obtain:

$$M_{0,j,k} = \int_0^1 x(1-F)^k\,dF, M_{1,j,0} = \int_0^1 xF^j\,dF \tag{12}$$

For parameter estimation of the frequency curves, the jth order probability weight moments $M_{i,j,0}$ are generally used. Only the first three orders of moments must be calculated because a three-parameter distribution line shape is typically used in general calculations.

Based on the discrete continuous series, compute its probability weight moments as:

$$M_0 = M_{1,0,0} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$M_1 = M_{1,1,0} = \frac{1}{n}\sum_{i=1}^{n} \frac{(n-i)}{(n-1)} x_i$$
$$M_2 = M_{1,2,0} = \frac{1}{n}\sum_{i=1}^{n} \frac{(n-i)(n-i-1)}{(n-1)(n-2)} x_i$$
$$M_3 = M_{1,3,0} = \frac{1}{n}\sum_{i=1}^{n} \frac{(n-i)(n-i-1)(n-i-2)}{(n-1)(n-2)(n-3)} x_i \quad (13)$$

The linear moments can thus be expressed as:

$$\lambda_1 = M_0$$
$$\lambda_2 = 2M_0 - M_1$$
$$\lambda_3 = 6M_0 - 6M_1 + M_2$$
$$\lambda_4 = 20M_0 - 30M_1 + 12M_2 - M_3 \quad (14)$$

Similar to the conventional moment, the statistical characteristic parameters of the linear moments are defined as:

$\tau = \frac{\lambda_2}{\lambda_1}$ is the linear coefficient of variation, $L - Cv$;

$\tau = \frac{\lambda_3}{\lambda_2}$ is the linear coefficient of skewness, $L - Cs$;

$\tau = \frac{\lambda_4}{\lambda_2}$ is the linear coefficient of kurtosis, $L - Ck$.

The linear moment method is generally used in developed countries such as the United Kingdom and United States. It has the advantages of unbiasedness, good robustness, and high accuracy, and can be widely used in the estimation of hydrological frequencies such as design floods and extreme rainfall.

D. Theoretical Comparison between Linear Moment Method and Conventional Moment Method

Statistics show that, in the conventional moment method, for the p-order origin moment of a random variable X, when p=1,

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}(X_i) = \mu \quad (15)$$

That is, the sample mean $\overline{X}$ is an unbiased estimate of the overall mean μ. However, when p > 1, $(\overline{X})^p$ is not an unbiased estimate of μ^p. For example, when p=2 and,

$$E(\bar{X}^2) = D(\bar{X}) + [E(\bar{X}^2)]^2 = \frac{\sigma^2}{n} + \mu^2 \neq \mu^2 \quad (16)$$

For a p-order central moment of random variable X, when p = 2,

$$\hat{\sigma}^2 = E(S^2) + E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$
$$= \frac{1}{n}E\left\{\left[\sum_{i=1}^{n}[(X_i - \mu) - (\bar{X} - \mu)]^2\right]\right\}$$
$$= \frac{1}{n}\left[\sum_{i=1}^{n}E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2\right]$$
$$= \frac{n-1}{n}\sigma^2 < \sigma^2 \quad (17)$$

It clear that the sample second-order central moment S2 is not an unbiased estimate of the overall variance σ2, while the estimator $\hat{\sigma}^2 < \sigma^2$. It can be shown that the p-order central moments of the samples are not unbiased estimates of the overall p-order central moments when p > 2 and that the degree to which the estimates are biased increases as p increases.

In a hydrological frequency analysis, it is often necessary to estimate Cv and Cs, which leads to hydrological frequency calculations and predictions.

If the conventional moment method is used for parameter estimation, second-, third-, or even higher-order moments of the origin are required. Therefore, Cs estimated by the conventional moment method is large, which can be proven to result in a small and unsafe estimate P of the design flood and heavy rainfall when applying the following Pearson type-III curve estimation formula. The Pearson type-III curve inverse function in terms of Cv and Cs can be expressed as:

$$P = \frac{\left(\frac{2}{\bar{X}C_v C_s}\right)^2}{\Gamma\left(\frac{4}{C_s{}^2}\right)} \int_X^{\infty} \left(X - \bar{X} + \frac{2C_v}{C_s}\bar{X}\right)^{\frac{4}{C_s{}^2}-1} \times exp\left[-\frac{2}{\bar{X}C_v C_s}\left(X - \bar{X} + \frac{2C_v}{C_s}\bar{X}\right)\right] dx \quad (18)$$

From the analysis provided, it is clear that the linear moment method demonstrates significantly greater robustness in parameter estimation than the conventional moment method. In the linear moment method, the moments of each order are linear combinations of the expected values of the sample-order statistics. In practical applications, only addition and subtraction operations using sample information are required, which significantly reduces errors that exist in the sample itself. Simultaneously, according to the definition of various linear moment coefficients, the corresponding linear moments are divided; however, the conventional moment method is also reduced in the calculation of bias coefficients used in the calculation of the third-order moments of the error generated. Therefore, the linear moment method is theoretically superior to the conventional methods. The following table 1 provides a list of comparisons between the conventional and linear moments.

Table 1: Comparison of Conventional Moments and Linear Moments

conventional moments		linear moments	
formulation	clarification	formulation	clarification
$\mu = E(X)$	first-order moment average value	$\mu = E(X_{1:1})$	one-dimensional combination average value
$\mu_2 = E(X-\mu)^2$ $\sigma = \mu_2^{1/2}$ $Cv = \left(\mu_2^{1/2}\right)/\mu$	Second-order moments (deviations) standard deviation devariation coefficient	$\mu = E(X_{1:1})$ $L - Cv = t = \lambda_2/\lambda_1$	binary combination linear deviation coefficient
$\mu_3 = E(X-\mu)^3$ $Cs = \mu_3/\left(\mu_2^{1/2}\right)^3$	third-order moments bias coefficient	$\lambda_3 = \frac{1}{3}E(X_{3:3} - 2X_{2:3} + X_{1:3})$ $L - Cs = t_3 = \lambda_3/\lambda_2$	ternary combination linear bias coefficient
$\mu_4 = E(X-\mu)^4$ $Ck = \mu_4/\left(\mu_2^{1/2}\right)^4$	fourth-order moments kurtosis coefficient	$\lambda_4 = \frac{1}{4}E\left(X_{4;4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}\right)$ $L - Ck = t_4 = \lambda_4/\lambda_2$	quadratic combination linear kurtosis coefficient

Comprehensive comparative analysis show that the linear moment method is the optimal algorithm among many current parameter estimation methods, and it is also most widely used algorithm in the international community. The linear moment method is less affected by the length of the data series [28-29], so it is widely used in the regional frequency analysis of rainstorms.

III. CASE STUDY - JIANGSU PROVINCE AS AN EXAMPLE

A. Overview of the Study Area

Jiangsu Province is located between 116°18′-121°57′ east longitude and 30°45′-35°20′ north latitude, situated in the Yangtze River Delta, with a flat terrain, vast plains, no steep mountains, numerous lakes, a dense water network, and borders the Yellow Sea. Jiangsu Province is located in a subtropical to warm temperate transitional climate zone. Its climate has obvious monsoon characteristics, with dry and cold winter, hot and humid summers, four distinct seasons, a mild climate, moderate rainfall, rain and heat in the same season, and sufficient light energy. The province's multi-year average rainfall is 700-1250 mm, and rainfall is abundant; however, the spatial and temporal distributions are uneven. In Jiangsu Province, the bulk of heavy rainfall occurs in the flood season, spanning four months from June through September. From the analysis of the causes of heavy rainfall and flooding, the early period (early June to early July) of the plum rains, late typhoon storms, and plum rains is prone to basin-wide flooding, and typhoons are typically the main cause of regional flooding. In addition, short-calendar-time thunderstorms are prone to small-scale localized flooding. Jiangsu Province, positioned within the Yangtze River Economic Belt, leads the nation in GDP per capita, regional development, and the Development and Livelihood Index (DLI), identifying it as one of the most holistically developed provinces in China. The swift advancement of the economy and urbanization has accelerated the improvement of the construction level of water conservation projects, and the requirements for hydrological frequency analysis are also increasing. Therefore, determining a better parameter estimation method that yields more accurate estimation results has become the focus of hydrological frequency analysis in Jiangsu Province.

B. Rainfall Frequency Calculation and Parameter Estimation

This research selected annual extreme value data from 149 long-duration representative rainfall stations in Jiangsu Province, covering four different durations: 1-day, 3-day, 7-day, and 15-day periods. The random forest

model was programmed using the Scikit-learn library in the Python language, employing long-series data as training samples to predict missing values in precipitation sequences. The results indicate that the predictions are reliable and can enhance the data series. Based on this, the rainfall frequency was analyzed using both conventional and linear moment methods. Figure 2 depicts the distribution of rain stations.

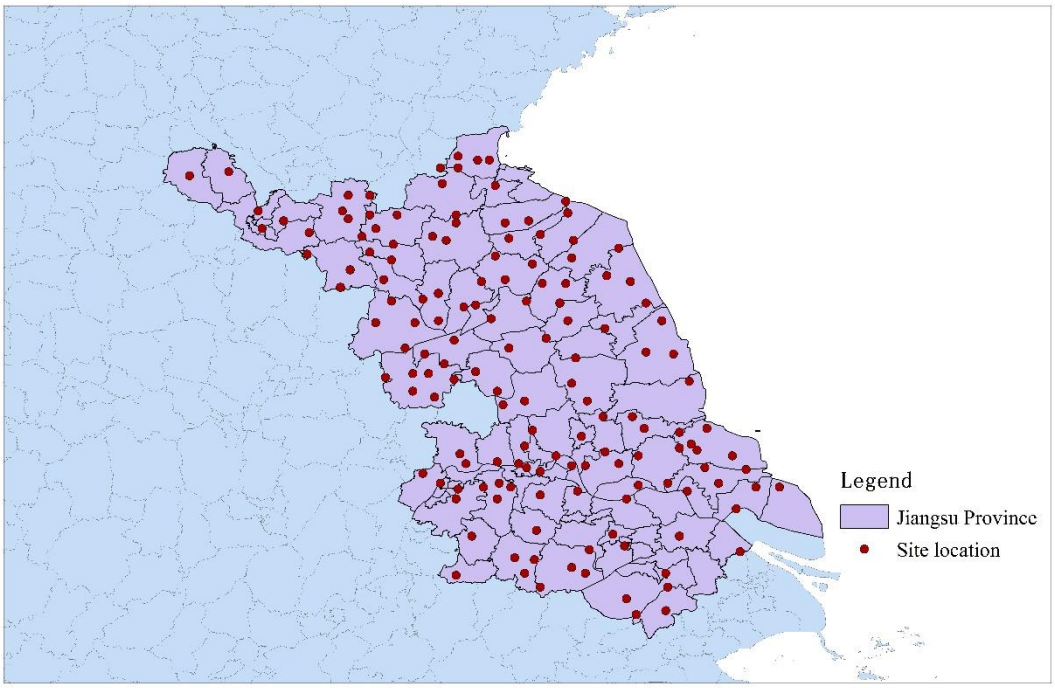


Figure 2: Distribution of Rainfall Stations in Jiangsu Province

The station storm frequency values were determined using the linear moment method. This method, through regional analysis, estimated the annual extreme rainfall frequency values for each station within the region for various recurrence periods by combining both local and general components. This method makes use of historical data of other stations in the region, fully considers information on the characteristics of local stations, and comprehensively analyzes the rainfall frequency distribution curves for each rainfall station in the region. The regional analysis assumes that the rainfall series at each site can be divided into two parts: a regional component reflecting rainfall characteristics common to the region and a local component reflecting rainfall characteristics specific to the region. Based on this assumption, the commonalities are partitioned regionally to obtain a dimensionless frequency distribution curve, while individuality is left local. The commonality and individuality are then superimposed to obtain a single-site frequency estimate. Its advantage is that it fully utilizes the historical information of other stations in the region and fully considers the characteristics of the local station information to carry out a comprehensive analysis of the rainfall frequency distribution curve of each rainfall station in the region and then deduces the precision and accuracy of each station with higher rainfall frequency estimates.

The method of calculating the frequency values of station storms using the conventional moment method is as follows: a preliminary estimation of different frequency values is carried out for different stations and time period series values, followed by fitting the estimated values for different stations and time periods using P-III type curves to derive design values for different frequencies.

According to the SL44-2006, the P-III distribution frequency curve was used to calculate the design flood, and its distribution function is as follows:

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}(x-\alpha_0)^{\alpha-1}e^{-\beta(x-\alpha_0)} \tag{19}$$

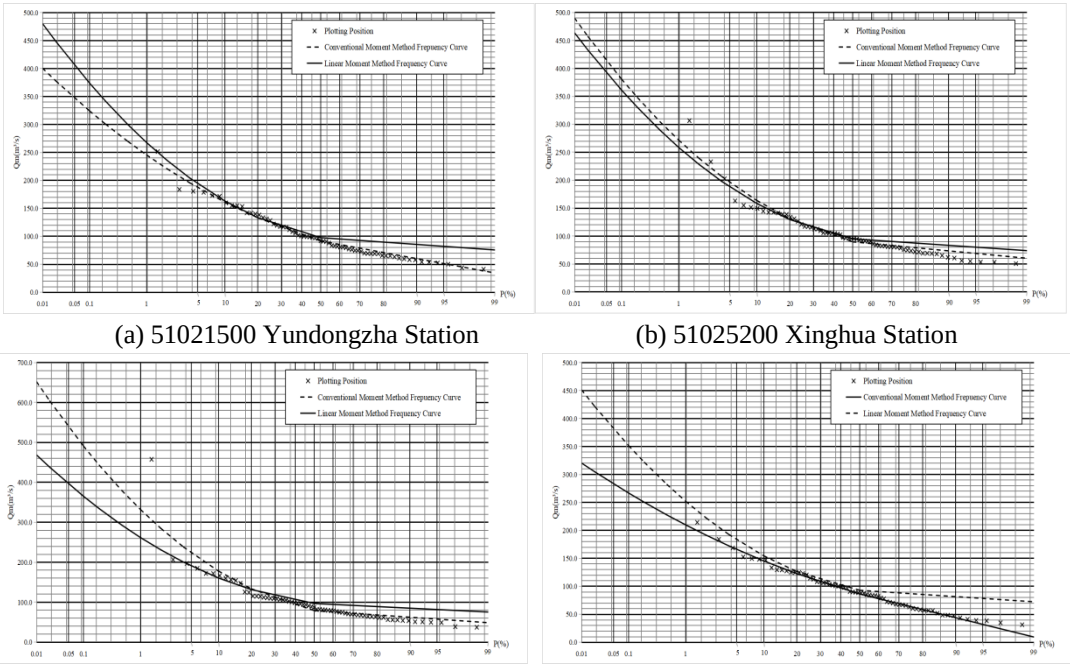
Where α, β, and α0 are the shape, scale, and location parameters of the distribution function respectively, and they are related to the three commonly used statistical parameters $\bar{x}$, Cv and Cs as follows:

$$\alpha = \frac{4}{C_s^2}, \alpha = \frac{2}{\bar{x}C_vC_s}, \alpha_0 = \bar{x}(1-\frac{2C_v}{C_s}) \tag{20}$$

The frequency curve obtained by the fit-line method fits better with the empirical data; thus, the fit-line method is primarily used to estimate the statistical parameters of the frequency curve in China [12].

To make a more intuitive and precise comparison of the differences between the rainfall frequency estimates obtained by the conventional and linear moment methods under different recurrence periods, four stations with

longer station sequences in the 1-day extreme rainfall event in Jiangsu Province were selected for comparison. Because the constant rule method plus the P-III curve is primarily used in China to fit rainfall or flood data to estimate its frequency estimates under different return periods, only the single-site P-III curve based on the constant rule method, and the P-III curve based on the regional linear moments method for rainfall estimation were compared here and plotted as a dot plot, as shown in Figure. 3.

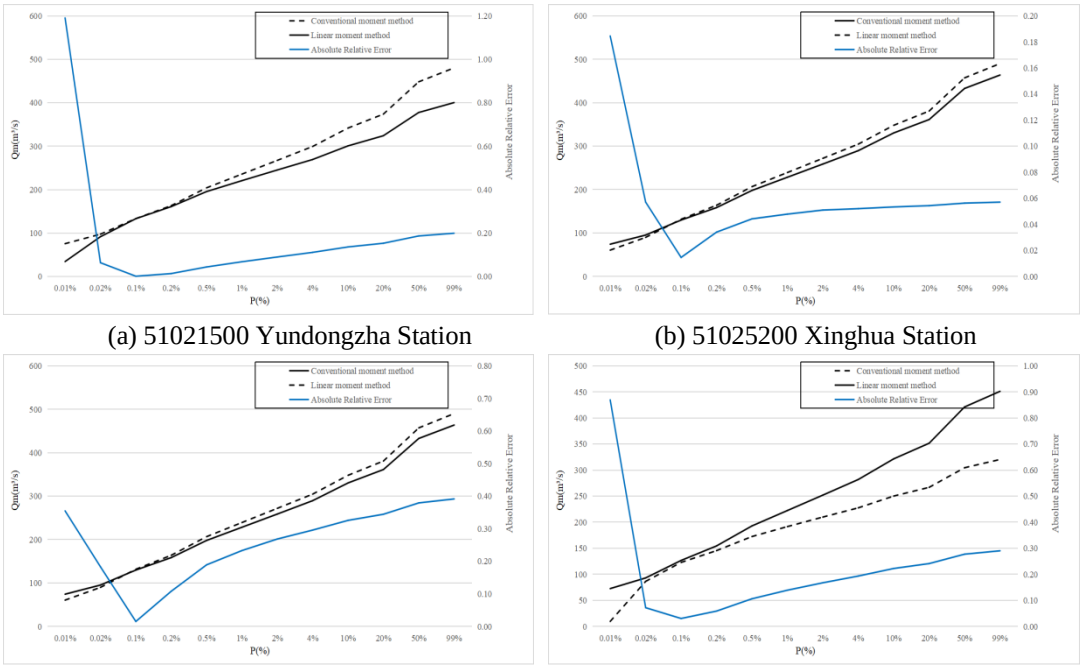


(a) 51021500 Yundongzha Station (b) 51025200 Xinghua Station

(c) 51029900 Yaowanggangzha Station (d) 51031100 Qinglonggang Station

Figure 3: Comparison of Rainfall Estimation Results between the Conventional Moment Method and the Linear Moment Method for P-III Curves

The relative error values of rainfall designed by linear and conventional moment methods under each recurrence period were calculated and compared, as shown in Figure. 4.



(a) 51021500 Yundongzha Station (b) 51025200 Xinghua Station

(c) 51029900 Yaowanggangzha Station (d) 51031100 Qinglonggang Station

Figure 4: Comparison of Rainfall and Relative Error Designed by Conventional Moment Method and the Linear Moment Method under Each Recurrence Period

Except, for the recurrence period of 1 year, if the design frequency is less than 10%, the recurrence period is less than 10 years, the disparity in the design rainfall determined by the conventional moment method versus the

linear moment method is minimal. As the design frequency rises, the disparity slowly widens, and the frequency estimates derived from the two methods show larger variances as the recurrence interval extends.

To analyze the reasons for the above differences, first of all, the conventional moment method lacks objective criteria, the curve and point data may not be completely fitted, the results are affected by human factors to a certain extent, and there is general subjective arbitrariness. Furthermore, the selection of statistical parameters takes into account incomplete factors, resulting in certain sampling errors. Second, when frequency calculations were performed, the lack of information on the historical return period for the occurrence of rainfall mega-values in past years prevented the wiring from adequately considering the maximum value of the rainfall series, and there were certain errors. As the length of the sample series utilized is only approximately 70 years, it is less reliable for estimating rainfall over a return period of more than 100 years. Therefore, when comparing the outcomes of the constant rule approach with the linear method of moments at individual stations, notable differences emerge for return periods exceeding 100 years.

C. Comparison of Estimated Parameters

1) Unbiasedness test

The unbiasedness test of the parameter estimation method adopted the ideal sample reduction method for a comparative study[30], where the sample size was the number of sites calculated in this frequency calculation, and the design frequency was P=1%. The formula is as follows:

$$X_{(1\%)} = E_x[1 + C_v\phi(P, C_s)] \tag{21}$$

Where Ex is the mean value and φ (P, Cs) is the dispersion coefficient. The correspondence between frequency P, skewness coefficient Cs and dispersion coefficient φ (P, Cs) has been made into a numerical table. For a given frequency P and a determined Cs, the corresponding deviation mean coefficient φ can be obtained. Table 2 displays the results of the unbiasedness test, highlighting only the outcomes and deviations for the mean values, Cv, Cs and P = 1% estimates for the four sites in the sample are listed. The values in the table are expressed as relative deviations (the percentage of deviation between the estimated value and the population value in the population value), and the absolute average values of the relative deviations of each sample mean value, Cv, Cs and P = 1% are listed in the table. The average deviation figures presented in Table 2 serve as an integrated index for assessing the unbiased characteristics of the proposed methodology. A smaller absolute average of the relative deviation indicates a higher degree of unbiasedness in the parameter estimation method. From the calculation results, the average deviations for the estimated values of Cv, Cs, and P = 1%, when computed using the linear moment method, are found to be lower than those derived through the conventional moment method. Thus, the parameters estimated by the linear moment method are more accurate than those of the conventional moment method, and it surpasses the conventional moment method regarding the unbiased nature of the estimation process.

Table 2: Unbiasedness Comparison Test Results of P-III Distributed Parameter Estimation Methods

				(%)
Station number	item	Theoretical value	Conventional moment method	Linear moment method
50437600	E_x	104.95	4.8	4.8
	C_v	0.44	9.3	0.2
	C_s	1.40	25.6	7.0
	$X_{(1\%)}$	254.27	17.6	8.6
50941400	E_x	109.17	6.7	6.7
	C_v	0.46	7.5	-3.2
	C_s	1.31	27.0	5.7
	$X_{(1\%)}$	272.36	8.8	-1.1
51025200	E_x	104.27	2.7	2.7
	C_v	0.41	7.3	2.4
	C_s	2.14	-2.3	-0.9
	$X_{(1\%)}$	259.48	4.7	-0.3
51027850	E_x	117.99	3.2	3.2
	C_v	0.45	8.3	-0.5
	C_s	1.14	35.5	4.0
	$X_{(1\%)}$	284.27	9.8	2.8
Mean deviation	E_x		4.0	4.0
	C_v		10.3	3.9
	C_s		29.8	4.0
	$X_{(1\%)}$		16.8	7.1

2) Robustness test

Robustness is an important index for measuring the stability of parameter estimation methods and reflects the degree of simulation of singular values. Simultaneously, the sample size has a certain impact on the precision of parameter estimation. In general, the larger the sample capacity, the smaller the influence of the singular value, the higher the precision of the parameter estimation, and the stronger its robustness. Conversely, the robustness of the parameter estimation method was inferior.

When testing the robustness of each parameter estimation method, we only consider whether they are significantly affected by individual point data (maximum or minimum). In the test, it was assumed that there were K extreme points in the actual sample that were contaminated; the peak value rose and the nadir value dropped while the recurrence period remained constant. For the other series of unchanged samples, the original method was used to estimate the statistical parameters and design values.

Comparing the deviation of the "contaminated" sample from the design value of the original sample, the smaller the deviation, the less the method is affected by individual points. Deviations were calculated using the following formula:

$$delx_{X_{p(i)}} = \frac{\left[X_p(i) - X_{p^0}(i)\right]}{X_{p^0}(i)} \times 100\% \tag{22}$$

$$Mx_p = \frac{1}{L}\sum_{i=0}^{n}\left|\frac{\left[X_p(i) - X_{p^0}(i)\right]}{X_{p^0}(i)}\right| \times 100\% \tag{23}$$

Where L is the number of samples taken and Xp (i) and Xp0(i) are the estimated value of the contaminated sample and the theoretical value of the sample when the sample number is i, respectively. The robustness of the parameter estimation method was tested using an ideal sample reduction method. The sample size was the number of sites calculated in this frequency calculation, and P was set to 4%, 1% and 0.1%. The robustness test results are shown in Table 3, where only the values P = 4%, P = 1% and P = 0.1% of the four locations with exceptionally high or low figures in the dataset are documented, and the values in the table are expressed in terms of relative deviation (the percentage of the deviation between the estimated value and the total value in the total value). The average deviation results in Table 3 can be used as a comprehensive index to evaluate the robustness of the proposed method. The smaller the absolute average of the relative deviation, the better the robustness of the parameter estimation method. The results show that the average deviation of P = 4%, P = 1% and P = 0.1% estimates calculated by the linear moment method is lower than the average deviation calculated by the conventional moment method; thus, the linear moment method has great advantages over the conventional moment method in terms of the robustness of the estimation process.

Table 3: Robustness Comparison Test Results of P-III Distributed Parameter Estimation Methods (%)

Station number	item	Theoretical value	Conventional moment method	Linear moment method
50939100	$X_{(4\%)}$	206.19	3.7	-1.2
	$X_{(1\%)}$	266.85	4.3	-0.8
	$X_{(0.1\%)}$	366.28	4.7	0.3
51023450	$X_{(4\%)}$	204.94	6.7	-5.8
	$X_{(1\%)}$	272.29	8.1	-7.4
	$X_{(0.1\%)}$	384.44	10.9	-8.5
51024600	$X_{(4\%)}$	211.57	5.6	-3.8
	$X_{(1\%)}$	283.28	6.7	-5.8
	$X_{(0.1\%)}$	402.30	7.8	-7.1
51026600	$X_{(4\%)}$	204.71	4.0	-1.5
	$X_{(1\%)}$	270.73	5.2	-2.6
	$X_{(0.1\%)}$	387.62	5.3	-5.0
Mean deviation	$X_{(4\%)}$		6.2	2.0
	$X_{(1\%)}$		8.1	2.1
	$X_{(0.1\%)}$		10.5	2.3

IV. CONCLUSIONS

In this study, the differences between the linear moment method and the conventional moment method in parameter estimation were compared from both theoretical and practical perspectives. The frequency analysis used the P-III distribution curve, which most commonly used in hydrological frequency analysis, and rainfall data from 149 rain stations in Jiangsu Province were processed using the random forest algorithm. The outcomes and

differences between the two parameter estimation methods were analyzed and compared, yielding the following conclusions and recommendations:

(1) In theory, the linear moment, as the expected value of the linear combination of sample-order statistics, significantly reduces the estimation error caused by using the samples' second- and third-order central moments in the normal method.

(2) Using rainfall data from Jiangsu Province as an example, simulations and predictions were made using the random forest model, which can also supplement missing rainfall data in practical applications. Based on this, frequency calculations were compared and analyzed using both the conventional moment and linear moment methods. The results show that when the design frequency is low, the error of the conventional moment and linear moment methods are small, and both can be used to calculate the design rainfall. The error of linear moment method was smaller and closer to the actual value as the design frequency increased.

(3) The study and discussion of the parameter estimation methods, particularly those based on the measured data, and the comparison of the linear moment method and the conventional moment method revealed that the linear moment method outperforms the conventional moment method in terms of the unbiasedness and robustness of the estimation process, and that the parameters estimated by the linear moment method are more accurate than those estimated by the conventional moment method.

The risk of floods triggered by extreme climate change poses even more severe challenges to society and physical infrastructure, which means that the precision and accuracy of hydrological frequency analysis needs to be pursued to a higher level. Currently, hydrologic frequency analysis in China is still based on the traditional method of artificial line fitting, which only focuses on linear distribution and parameter estimation in the case of a single station, and has significant limitations. For example, the single-station and single-period analysis method is heavily influenced by errors, and the design value estimated by the normal method is obviously small. Furthermore, the line-fit method necessitates a very high level of calibration and experience from the worker; therefore, its estimated design values are typically poor in terms of precision and accuracy. Although the P-III distribution curve is the standard requirement for calculating hydrological frequency in China, it may not be the optimal linear distribution in a given region and is therefore unsuitable for hydrological frequency test analysis. Simultaneously, frequency curves and corresponding statistical parameters used in engineering should be analyzed not only in terms of hydrological statistics, but also in the context of the physical causes of hydrological phenomena and regional patterns. The linear moment method is useful not only for single-station hydrologic frequency calculations, but also for regional synthesis and line identification in undocumented areas. To reduce the error generated by the conventional moment method, frequency calculations can be performed using the optimized fit-line method, which optimizes the fit of empirical frequency data to the known theoretical frequency curve by establishing an objective function.

ACKNOWLEDGMENT

This work was partially funded and supported by the Major Scientific and Technological Projects of the Ministry of Water Resources (Granted: SKS-2022021), the Fundamental Research Funds for the Central Universities (Granted: B220202028), and the Jiangsu Provincial Water Resources Technology Project (Granted: 2017006). The authors also would like to thank the editor and reviewers for their crucial comments.

REFERENCES

[1] I. Leščešen, M. Šraj, B. Basarin, D. Pavić, M. Mesaroš, M. Mudelsee, Regional Flood Frequency Analysis of the Sava River in South-Eastern Europe, Sustainability 14 (2022) 9282.

[2] E. Dodangeh, V.P. Singh, B.T. Pham, J. Yin, G. Yang, A. Mosavi, Flood Frequency Analysis of Interconnected Rivers by Copulas, Water Resour. Manag. 34 (2020) 3533–3549.

[3] Z. Li, F. Brissette, J. Chen, Assessing the applicability of six precipitation probability distribution models on the Loess Plateau of China, Int. J. Climatol. 34 (2014) 462–471.

[4] X. Tang, C. Miao, Y. Xi, Q. Duan, X. Lei, H. Li, Analysis of precipitation characteristics on the loess plateau between 1965 and 2014, based on high-density gauge observations, Atmospheric Res. 213 (2018) 264–274.

[5] L. Yin, F. Tao, Y. Chen, F. Liu, J. Hu, Improving terrestrial evapotranspiration estimation across China during 2000–2018 with machine learning methods, J. Hydrol. 600 (2021) 126538.

[6] M.K. Tiwari, C. Chatterjee, Uncertainty assessment and ensemble flood forecasting using bootstrap based artificial neural networks (BANNs), J. Hydrol. 382 (2010) 20–33.

[7] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32.

[8] H. Park, K. Kim, D. kun Lee, Prediction of Severe Drought Area Based on Random Forest: Using Satellite Image and Topography Data, Water 11 (2019) 705.

[9] R. Majumder, B.J. Reich, A deep learning synthetic likelihood approximation of a non-stationary spatial model for extreme streamflow forecasting, Spat. Stat. 55 (2023) 100755.
[10] A. Castellarin, D.H. Burn, A. Brath, Assessing the effectiveness of hydrological similarity measures for ¯ood frequency analysis, J. Hydrol. (2001).
[11] V.P. Singh, W.G. Strupczewski, On the status of flood frequency analysis, Hydrol. Process. 16 (2002) 3737–3740.
[12] Jin, G. 1999. A Review of Hydrologic Frequency Analysis. ADVANCES IN WATER SCIENCE. 10 (3):319-327.
[13] S. Benameur, A. Benkhaled, D. Meraghni, F. Chebana, A. Necir, Complete flood frequency analysis in Abiod watershed, Biskra (Algeria), Nat. Hazards 86 (2017) 519–534.
[14] A.S. Kebebew, A.A. Awass, Regionalization of catchments for flood frequency analysis for data scarce Rift Valley Lakes Basin, Ethiopia, J. Hydrol. Reg. Stud. 43 (2022) 101187.
[15] Á. Ossandón, B. Rajagopalan, W. Kleiber, Forecasting Magnitude and Frequency of Seasonal Streamflow Extremes Using a Bayesian Hierarchical Framework, Water Resour. Res. 59 (2023) e2022WR033194.
[16] J.A. Greenwood, J.M. Landwehr, N.C. Matalas, J.R. Wallis, Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form, Water Resour. Res. 15 (1979) 1049–1054.
[17] R.K. Jaiswal, T.R. Nayak, A.K. Lohani, R.V. Galkate, Regional flood frequency modeling for a large basin in India, Nat. Hazards 111 (2022) 1845–1861.
[18] A.K. Mishra, V.P. Singh, Drought modeling – A review, J. Hydrol. 403 (2011) 157–175.
[19] J.R.M. Hosking, L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics, J. R. Stat. Soc. Ser. B Methodol. 52 (1990) 105–124.
[20] J.R.M. Hosking, Some theory and practical uses of trimmed L-moments, J. Stat. Plan. Inference 137 (2007) 3024–3039.
[21] A. Sankarasubramanian, K. Srinivasan, Investigation and comparison of sampling properties of L-moments and conventional moments, J. Hydrol. 218 (1999) 13–34.
[22] H.D. Fill, J.R. Stedinger, Homogeneity tests based upon Gumbel distribution and a critical appraisal of Dalrymple's test, J. Hydrol. 166 (1995) 81–105.
[23] Z. Hussain, Application of the Regional Flood Frequency Analysis to the Upper and Lower Basins of the Indus River, Pakistan, Water Resour. Manag. 25 (2011) 2797–2822.
[24] Liang. Y., Liu. S., Zhong. G., Zhou. Z., Hu. Y., 2013. Comparison between Conventional Moments and L-moments in Rainfall Frequency Analysis for Taihu Lake Basin. Journal of China Hydrology 33(4):16-21.
[25] C. Ilinca, C.G. Anghel, Flood Frequency Analysis Using the Gamma Family Probability Distributions, Water 15 (2023) 1389.
[26] B. Efron, Bootstrap Methods: Another Look at the Jackknife, in: S. Kotz, N.L. Johnson (Eds.), Breakthr. Stat. Methodol. Distrib, Springer, New York, NY, 1992: pp. 569–593.
[27] T.Zhao, D.Yang, X.Cai, Y,Cao, Predict seasonal low flows in the upper Yangtze River using random forests mode Abstract: Predicitor selecion and model construcion are two kev isues in lono-temm streamiowforecasina.This sudv inroduces a random foresis model for selecina, J. Hydroelectr. Eng. (2012) 18–24, 38.
[28] N. Papukdee, J.-S. Park, P. Busababodhin, Penalized likelihood approach for the four-parameter kappa distribution, J. Appl. Stat. 49 (2022) 1559–1573.
[29] Y. Shin, P. Busababodhin, J.-S. Park, The r-largest four parameter kappa distribution, (2020).
[30] Liu, G. 1990. Pearson type-III distribution parameter estimation. Journal of China Hydrology. (4-5):1-15.