The effect of time series length on runoff characteristics analysis

S Zhang¹, D F Liu¹, Q Huang¹, X Wei¹, J L Feng² and Mu Lin³

¹State Key Laboratory Base of Eco-hydraulic in Arid Area, School of Water Resources and Hydropower, Xi’an University of Technology, Xi’an 710048, China
²Shanxi Provincal Water and Soil Conservation and Ecological Environment Construction Center, Taiyuan Shanxi 030002, China
³School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 100081, China

E-mail: liudf@xaut.edu.cn

Abstract. The length of time series is of great importance, which is an uncertain factor in runoff characteristics analysis, especially in the analysis of trend and period. This paper investigates the tendency and period using runoff observation data of 107 years from 1909 to 2015 at Monticello hydrological station in the Upper Sangamon River in USA. The data series of 30 years, 40 years, 50 years and so on until 100 years and 107 years are selected to test their trend and period, respectively, and then are compared with each other. The results show that, with the increase of time series length, the tendency and period tend to be stable, and the annual runoff begins to show a significant increase trend once the length of time series is more than 80 years. The period tends to a dynamic stability value of 22 years when the time series length is more than 70 years. When the series length is larger than 70 years, the results of runoff characteristics analysis are much more reliable.

1. Introduction

In recent years, most of river runoff has changed to a great extent due to the impact of climate change and human activities [1]. Hence, runoff characteristics have been widely analysed to provide a basis for the rational development and utilization of water resources [2]. In the last few decades, scholars have been much more mature on the research of trend and period analysis of hydrological series. Sang et al [3] analysed the determined period by using empirical mode decomposition (EMD) and maximum entropy spectral analysis (MESA) which can efficiently avoid the impacts of noise on trend and period analysis. Burn and Hag Elnur [4] utilized the Mann-Kendall non-parametric test to detect trends and found out that a greater number of trends are observed than are expected to occur by chance. Hamed [5] improved Mann-Kendall test to study the significant trends in annual maximum flow of a group of 57 worldwide total annual runoff time series. Dixon et al [6] used the non-parametric Mann-Kendall test with bootstrap resampling to calculate annual and seasonal trends and then concluded that a marked east-west gradient of streamflow trend emerged in the mountainous west. Labat [7] gave a review of concepts of Wavelet analyses and the recent advances, which can be of great importance to identify period components. Gaucherel [8] exploited the potential of continuous wavelet transform on flow curves to detect new periodicities or time annual features. Sang et al [9] developed a discrete wavelet spectrum (DWS) approach to identify non-monotonic trend patterns in hydrological time
series and evaluate their statistical significance.

The length of time series plays an important role in hydrological research. Xu and Xie [10] established the Autoregressive Integrated Moving Average (ARIMA) model to analyse the influence of the length of time series for the results of monthly runoff forecast. However, up to now, few people have investigated the effects of time series length on the analysis of runoff characteristics. Therefore, in this paper, we preliminarily discuss the optimum length of time series using runoff observation data of 107 years from 1909 to 2015 at Monticello hydrological station in the Upper Sangamon River in USA.

2. Study area and data
The Upper Sangamon River Basin (USRB) is located in the Upper Illinois River in Illinois, USA, with a watershed area of about 3150 km². The United States Geological Survey (USGS) sets up a water level gage at Monticello (N40°01′51″, W88°35′20″) as shown in figure 1, with a catchment area of about 1430 km². Monticello water level gage has collected daily runoff observation of 107 years from 1909 to 2015, which provides important basic data for analysing the long-time characteristics of runoff in this watershed [11].

![Figure 1. Location of the Monticello hydrological station in the USRB.](image)

USRB has been affected by human activities including construction of railway and drainage systems, excavation of drains, laying drainage pipe underground and other activities since the 19th century. In the midstream of USRB, there is a manmade lake named Decatur, where a reservoir lies. Because of the lack of reservoir’s water supply data, the water level gage is finally selected to make runoff characteristics analysis, which is located in the upstream of Decatur Lake [12].

The runoff data are obtained from USGS. They are divided into 9 groups with a fixed year of 1909 and a step of 10 years. In other words, we take the runoff observation of 30 years from 1909 to 1938, 40 years from 1909 to 1948, …, 100 years from 1909 to 2008 and 107 years from 1909 to 2015. And
then their tendency and period are analyzed, respectively and are compared.

3. Methods

In this paper, three methods are applied to analyze the runoff characteristics, among which Correlation Coefficient Analysis (CCA) and Linear Regression Trend Test (LRTT) for tendency analysis and Wavelet analysis (WA) for period analysis.

3.1. Correlation coefficient analysis (CCA) and linear regression trend test (LRTT)

It is assumed that there is a linear trend component in a hydrologic sequence \( x_t \), and the formula is as follows:

\[
x_t = T_t + \eta_t = a + bt + \eta_t
\]

(1)

where, \( T_t \) represents the trend component; \( \eta_t \) represents other components; \( a \) and \( b \) represents the parameters of the linear equation in the trend component respectively; \( t \) is time. The expressions of each parameter are as follows [13]:

\[
b = \frac{\sum_{t=1}^{n} (t - \tilde{t})(x_t - \bar{x})}{\sum_{t=1}^{n} (t - \tilde{t})^2}
\]

(2)

\[
a = \bar{x} - b\tilde{t}
\]

(3)

\[
S_b^2 = \frac{S^2}{\sum_{t=1}^{n} (t - \tilde{t})^2}, \quad S^2 = \frac{\sum_{t=1}^{n} \eta_t^2}{n - 2}
\]

(4)

\[
\sum_{t=1}^{n} \eta_t^2 = \sum_{t=1}^{n} (x_t - \bar{x})^2 - b^2 \sum_{t=1}^{n} (t - \tilde{t})^2
\]

(5)

where, \( S_b \) represents the standard deviation of \( b \); \( \sum_{t=1}^{n} \eta_t^2 \) represents the sum of square of deviation; \( \bar{x} \) and \( \tilde{t} \) represent the mean of \( x_t \) and \( t \), respectively; \( n \) represents the capacity of sample.

\( r \) is taken as the correlation coefficient between \( x_t \) and \( t \), and \( T \) is defined as the statistic value, there calculation formulas are as follows:

\[
r = \frac{\sum_{t=1}^{n} (x_t - \bar{x})(t - \tilde{t})}{\sqrt{\sum_{t=1}^{n} (x_t - \bar{x})^2 \cdot \sum_{t=1}^{n} (t - \tilde{t})^2}}
\]

(6)

\[
T = b / S_b
\]

(7)

In this paper, there are two steps for trend analysis:

- \( r \) is calculated first, and then find the correlation coefficient checklist to identify whether the relationship between \( x_t \) and \( t \) is significant or not. In other words, we want to find out whether the time series have a specific change or not with the influence of time. If it is, continue to the second step to obtain what change it is; otherwise, stop it to get a conclusion.
that there is no relationship between $x_t$ and $t$, hence there is no tendency there.

- $T$ is calculated using equation (7). At the beginning, $b$ is assumed to be equal to 0, and $T$ is subject to t-distribution with $D = n - 2$. Giving $\alpha$, $t_{\frac{\alpha}{2}}$ can be checked in the checklist of $t$.

Then compare $|T|$ and $t_{\frac{\alpha}{2}}$, if $|T| > t_{\frac{\alpha}{2}}$, the initial assumption of $b$ will be rejected, that is to say that the trend component is significant; otherwise, the trend component can be ignored.

3.2. Wavelet analysis (WA)

WA, called the ‘mathematic microscope’ [14], is an important break-through in the development history of Fourier transform (FT), which can identify very weak period signal compared to FT and has a powerful local time-frequency resolution for description of local signal characteristics [15]. In this paper, WA is used to test period component of different length of time series, it is as follows [11]:

$$W_f(a) = \int_{-\infty}^{\infty} |W_f(a,b)|^2 db$$  \hspace{1cm} (8)

where, $W_f(a)$ represents wavelet variance; $a$ represents scale parameter; $b$ represents translation parameter; $W_f(a,b)$ represents wavelet coefficient.

Using WA, the figure of wavelet variance is shown to identify the period change of different time scales with the development of distribution of $a$. The peak in each figure is defined to be the major period of different time scales.

4. Results

4.1. Tendency analysis

| $n$/years | 30  | 40  | 50  | 60  | 70  | 80  | 90  | 100 | 107 |
|-----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $r$       | 0.138 | 0.0964 | 0.0173 | 0.0387 | 0.0721 | 0.1204 | 0.2156 | 0.208 | 0.2287 |
| $r_{0.05}$ | 0.361 | 0.312 | 0.270 | 0.254 | 0.235 | 0.220 | 0.207 | 0.197 | 0.190 |
| $r_{0.01}$ | 0.463 | 0.403 | 0.361 | 0.330 | 0.306 | 0.286 | 0.270 | 0.256 | 0.248 |

* $a$ The correlation coefficient of each sequence length when the probability level is 5%.
* $b$ The correlation coefficient of each sequence length when the probability level is 1%.

![Figure 2](attachment:correlation_coefficient.png)

**Figure 2.** Comparison of correlation coefficient.
Through CCA, the correlation coefficient of each group is calculated in table 1. Compared with the critical value of the correlation coefficient from the correlation coefficient checklist, i.e. $r_{0.05}$ and $r_{0.01}$, the results are shown in table 1 and figure 2. The correlation coefficient is between $r_{0.05}$ and $r_{0.01}$ when n is more than 80 years. In other words, when the length of hydrological sequence exceeds 80 years, it exhibits a specific change with the influence of time and the trend component begins to appear. Then we adopt LRTT to test the significance of tendency for the hydrological sequence whose length is more than 80 years. The statistic value $T$ is calculated using equation (7), and listed in table 2 along with the t-test value $t_{\%}$ where $\alpha$ is defined to be equal to 0.05. When n is greater than 80 years, $|T|$ is always larger than $t_{\%}$. So the increasing tendency is significant. Hence, when the length of time series is more than 80 years, the effect of trend becomes significant.

| n /years | 90     | 100    | 107    |
|----------|--------|--------|--------|
| $T$      | 93.15  | 109.86 | 138.05 |
| $t_{\%}$ ($\alpha =0.05$) | 1.987  | 1.984  | 1.98   |

4.2. Period analysis

The curves of wavelet variance have been shown in figure 3. It is obvious that the wavelet variance curves of each group are similar in shape. They all have 2 or 3 peak values, and the first, second, third peak value represents the first, second, third main period respectively [16]. And the results are shown in table 3 and figure 4. From figure 3 and table 3, the first main period is generally expressed as a half of the length of time series, which has a great possibility to be the pseudo period. The first and second main period is approximately equal to the multiple of the third main period, and therefore the third main period can be ignored. Examining the second main period, we can conclude that the period of USRB tends to a dynamic stability when the length of hydrological sequence is more than or equal to 80 years. Therefore, we recommend that the length of hydrological sequence should be more than or equal to 80 years when analyzing period component in order to get a better result.

(a) 30 years

(b) 40 years

(c) 50 years

(d) 60 years
Figure 3. The curves of wavelet variance of different groups.

Figure 4. The periods and the length of the data
Table 3. Comparison chart of different main period in each group.

|                | n /years |
|----------------|----------|
|                | 30       | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 107 |
| First main period | 11 | 14 | 24 | 31 | 38 | 42 | 49 | 54 | 58 |
| Second main period | 3 | 7 | 12 | 15 | 20 | 24 | 19 | 22 | 24 |
| Third main period | 3 | 6 | 7 | 13 | 10 | 9 | 10 | 15 |

5. Conclusion
Most of the runoff observation data of the watershed in China is about 60 years, which limits our research on runoff characteristics analysis. This paper discusses trend and period components using runoff observation data of 107 years from 1909 to 2015 at Monticello hydrological station in the Upper Sangamon River in USA, which are divided into 9 groups.

The results show that, in USRB, we can detect a significant increasing tendency when the length of series is more than 80 years; based on WA, and the curves of wavelet variance are used to identify the periods of different group. We find out that the length of series should be more than or equal to 80 years when analyzing period component in order to get a better result. Therefore, it is recommended that the length of hydrological series should not be less than 80 years when analyzing runoff characteristics. But future runoff can be completely different, that should be taken into account with the actual situation. This study can provide the reference for future research.

Acknowledgments
This work was partially supported by the National Natural Science Foundation of China (NSFC) (Grant No. 51309188, 51609270). Funding was also provided by the Basic Research Plan of Natural Science of Shaanxi Province (2016JQ5105), SRFDP (20136118120021) and National Department Public Benefit Research Foundation of Ministry of Water Resources (201501058).

References
[1] Huang S Z, Huang Q, Wang Y M and Chen Y T 2014 The study on the runoff change based on the heuristic segmentation algorithm and approximate entropy ASNUS 53 154-60
[2] Huang J X, Xu Z X and Gong T L 2007 Characteristics and driving factors of the runoff variations in the Yarlung Zangbo river J. China Hydrol. 27 31-5
[3] Sang Y F, Wang Z G and Liu C M 2012 Period identification in hydrologic time series using empirical mode decomposition and maximum entropy spectral analysis J. Hydrol. 424-425 154-64
[4] Burn D H and Hag Elnur M A 2002 Detection of hydrologic trends and variability J. Hydrol. 255 107-22
[5] Hamed K H 2008 Trend detection in hydrologic data: The Mann-Kendall trend test under the scaling hypothesis J. Hydrol. 349 350-63
[6] Dixon H, Lawler D M and Shamseldin A Y 2006 Streamflow trends in western Britain Geophys. Res. Lett. 33 L19406
[7] Labat D 2005 Recent advances in wavelet analyses: Part 1. A review of concepts J. Hydrol. 314 275-88
[8] Gauch erel C 2002 Use of wavelet transform for temporal characterisation of remote watersheds J. Hydrol. 269 101-21
[9] Sang Y F, Sun F B, Vijay P S, Xie P and Sun J 2017 A discrete wavelet spectrum approach to identifying non-monotonic trend pattern of hydroclimate data Hydrol. Earth Syst. Sci . Discuss. doi:10.5194/hess-2017-6
[10] Xu M and Xie Q Q 2014 Analysis of the effect of time series length on monthly runoff forecast based on ARIMA model J. Yangtze Univ. (Nat. Sci. Edit.) 11 6-10
[11] Liu D F, Lei J Q, Huang Q and Lin M 2014 Analysis of runoff evolution characteristics with
observation data for nearly hundred-year J. Water Resour. Water Eng. 25 24-8
[12] Liu D F, Tian F Q and Hu H P 2009 Sediment simulation at Upper Sangamon River Basin using THREW model IAHS Publication 335 187-95
[13] Huang Q and Zhao X H 2008 Theory and Method of Time Series Analysis and Forecast for River Runoff (Zhengzhou, Henan Province: The Yellow River Water Conservancy Press) 18-9
[14] Xie Y Y, Huang Q, Chang J X, Liu S Y and Wang Y M 2016 Period analysis of hydrologic series through moving-window correlation analysis method J. Hydrol. 538 278-92
[15] Yang Y, Shu H P, Ma J Z and Yang Z F 2017 Small and medium scale of climate variation characteristics of Baiyin city, Gansu Province J. Arid Land Resour. Environ. 31 126-31
[16] Shao X M, Xu Y Q and Yan C R 2006 Wavelet analysis of rainfall variation in the Yellow River Basin ASNUP 42 503-9