Check for updates

Water Resources Research

RESEARCH ARTICLE
10.1029/2021WR031606

Key Points:

- Correlation coefficient-based information criterion (CCIC) was proposed to determine the optimal order of regression models for quantifying dependence characteristics in hydrologic time series
- MC experiments verified higher and more stable accuracy of determining the true model order by CCIC than by Akaike Information Criterion and Bayesian Information Criterion used commonly
- Mean value of a time series had a big impact on the accuracy of CCIC for determining optimal order for regression models

Correspondence to:

Y.-F. Sang and Y. Li,
sangyf@igsnrr.ac.cn;
liyaqing@whu.edu.cn

Citation:

Xie, P., Huo, J., Sang, Y.-F., Li, Y., Chen, J., Wu, Z., & Singh, V. P. (2022). Correlation coefficient-based information criterion for quantification of dependence characteristics in hydrological time series. *Water Resources Research*, *58*, e2021WR031606. https://doi.org/10.1029/2021WR031606

Received 11 NOV 2021
Accepted 7 JUL 2022

Author Contributions:

Conceptualization: Ping Xie, Jie Chen
Formal analysis: Jingqun Huo, Yaqing Li, Ziyi Wu
Investigation: Ping Xie, Jingqun Huo, Yan-Fang Sang
Methodology: Ping Xie, Jingqun Huo, Yan-Fang Sang, Yaqing Li, Ziyi Wu
Supervision: Ping Xie, Yan-Fang Sang, Vijay P. Singh
Validation: Ping Xie, Yan-Fang Sang, Yaqing Li, Ziyi Wu
Writing – original draft: Ping Xie, Jingqun Huo, Yaqing Li
Writing – review & editing: Yan-Fang Sang, Jie Chen, Vijay P. Singh

© 2022. American Geophysical Union. All Rights Reserved.

Correlation Coefficient-Based Information Criterion for Quantification of Dependence Characteristics in Hydrological Time Series

Ping Xie[1], Jingqun Huo[1], Yan-Fang Sang[2,3], Yaqing Li[4], Jie Chen[1], Ziyi Wu[1], and Vijay P. Singh[5,6]

[1]State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan, China, [2]Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China, [3]Key Laboratory of Compound and Chained Natural Hazards, Ministry of Emergency Management of China, Beijing, China, [4]Yangtze Ecology and Environment Co Ltd, Wuhan, China, [5]Department of Biological and Agricultural Engineering and Zachry Department of Civil & Environmental Engineering, Texas A&M University, College Station, TX, USA, [6]National Water and Energy Center, UAE University, Al Ain, UAE

Abstract Quantification of the dependence characteristics in hydrological time series is essential for understanding hydrological variability and for managing water resources. However, how determining a suitable model for describing the dependent components is still a challenge. In this article, we proposed a correlation coefficient-based information criterion (CCIC) to determine the optimal model order of regression-based models for quantifying the dependence characteristics in hydrological time series. CCIC was developed by combining the index of correlation coefficient and the information entropy index. The former was used to reflect the fitting error of the model and the latter was used as a penalty term to reflect the model uncertainty. They have similar roles as the two terms of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) used commonly in hydrology. Results of Monte-Carlo experiments verified higher and more stable accuracy of determining the true model order by CCIC than by AIC and BIC. Moreover, results indicated that the mean value of a time series had a big impact on the accuracy of CCIC. If the mean value of a time series was far from its initial value, the estimation of CCIC would have a big bias, causing low accuracy in the determination of suitable model order. The application of CCIC to annual precipitation from 520 stations in China further confirmed its advantages over AIC and BIC and indicated the significant short-term dependence characteristics of annual precipitation in the Yangtze River basin. The proposed CCIC approach has potential for wider use in hydrometeorology.

1. Introduction

Hydrological time series are analyzed for investigating the variability of hydrological processes (Machiwal & Jha, 2012). Apart from deterministic characteristics, including abrupt changes, trends, and periodicities (Sang et al., 2015; Xie et al., 2019; Zhang et al., 2014), dependence (also called persistence) is another important intrinsic characteristic of hydrological processes (Bomblies et al., 2008; Iliopoulou et al., 2018; Markonis et al., 2018). Generally, dependence means that the value of a hydrological variable at a certain time is not random but related to its previous values (Moravej & Khalili, 2015). For example, drought or flood events usually have a propensity to occur in clusters (Moravej & Khalili, 2015; Paschalis et al., 2012; Tan et al., 2017). Determining the dependence characteristics of hydrological processes is necessary to understand their complex variations and for hydrological calculation and design (Debele et al., 2017; Jiang et al., 2015; Zhao et al., 2018), simulation, and prediction (Ye et al., 2018), as well as water management and planning (Hodgkins et al., 2017; Koirala et al., 2011).

Both short- and long-term dependencies are investigated for hydrological time series analysis. For a stationary hydrological time series, the methods to detect its short-term dependence are mainly divided into non-parametric methods and parametric methods. Typical non-parametric methods include Spearman rank order serial correlation test (Zar, 1972) and rank von Neumann ratio (Gaonkar et al., 2021). Parametric methods, such as the autocorrelation coefficient test (Fathian et al., 2016), are commonly used for detecting short-term dependence. The Hurst coefficient is a dominant measurement available for quantifying long-term dependence (Markonis et al., 2018). In this study, we focus on the short-term dependence of hydrological processes.

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The Auto-Regressive (AR) models and Moving Average (MA) models, as well as their combination of Auto-regressive Moving Average (ARMA) models, are one important type of regression-based models (Hossain et al., 2020). They have been widely used to intuitively describe the short-term dependence within a time series (Markonis et al., 2018) and describe the statistical characteristics of residual series after removing the deterministic components in the original time series, due to their inherent computational efficiency (Yue & Pilon, 2003). Regarding the applications of this type of model, the key issues are to determine the suitable model order and estimate parameters (Han et al., 2017; Moon et al., 2021; Peng et al., 2009). The methods commonly used to estimate parameters include the least square estimation, moment estimation (Yule-Walker equations), and maximum likelihood estimation. However, the more important issue in using these models is the determination of suitable model order. The efficiencies of these models critically depend on the selection of appropriate model order, as lower model orders would provide inadequate information, while higher model orders could drastically increase complexity and cause model overfitting problems (Khan et al., 2021; Khorshidi et al., 2011).

The methods used to determine the optimal model order of regression-based models can be roughly divided into two categories: information criterion-based methods and linear algebra methods. The former mainly include Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC), which have been used widely in hydrology; the latter include Singular Value Decomposition (SVD) method, Levinson-Durbin methods, etc. Sometimes we also directly consider the statistical characteristics of the autocorrelation function and (or) partial correlation function of a time series for determining the suitable model order. When handling small data samples, AIC tends to select more complex models with a larger model order, due to a lower degree of punishment (Zhang, 2007). The accuracy of BIC in the determination of the optimal model order is always affected by the number of data samples (Lin et al., 2017). Comparatively, the SVD method is less sensitive to the data length and can be applied to shorter time series; however, it needs to solve a high-order algebraic equation for obtaining eigenvalues, which would significantly increase the complexity of calculation, and there may be a large bias when judging small model orders (Fort et al., 1995). For the Levinson-Durbin method, it is still based on AIC and BIC for the determination of the suitable model order, by first applying the Levinson-Durbin algorithm for parameter estimation (Franke, 1985; Liu et al., 2012). Thus, it is still a challenge to accurately determine a suitable model order for describing the short-term dependence of hydrological time series by using these regression-based models.

Generally, the determination of a suitable model order is directly determined by the dependent component of the time series to be analyzed. The essence of identifying the dependent component is to evaluate its statistical significance. In practical applications, the lag-1 autoregressive coefficient was often used to quantify the dependence of a time series (Sagarika et al., 2014). Comparing the estimated lag-1 autoregressive coefficient with the threshold value at a certain significance level, it can be judged whether the dependence exists (i.e., passing the significance test) or not (Serinaldi & Kilsby, 2016). However, this approach neglects the dependence with higher orders, such as that described by lag-2 and lag-3 AR models. Furthermore, it cannot classify the different significance levels of the dependence characteristics, which causes difficulty in linking the identification of dependent components to the determination of the suitable model order for AR, MA, or ARMA models.

Considering that the dependent component is part of the original time series, we can quantify the correlation between the dependent component and the original time series, which can be an effective approach to quantifying the dependence characteristics. Based on this idea, the objective of this study is to propose a new method, called correlation coefficient-based information criterion (CCIC), to evaluate the dependence characteristics of a hydrological time series and further determine the suitable model order for describing it. Combining the effective index of correlation coefficient (CC) and the information criterion, CCIC has a similar form as AIC and BIC. A bigger CC value usually represents a larger proportion of the dependent component in the original time series. The advantage of CCIC is to simultaneously identify the dependent component, assess the significant degree of dependence, and further determine the suitable model in order to describe it using AR, MA, or ARMA models. To this end, Section 2 explains the relationship between the correlation coefficient and serial dependence, based on which the CCIC is developed. In Section 3, different types of synthetic data are generated by the Monte-Carlo method, which is used to verify the efficacy of the proposed CCIC, and further evaluate the influence of main factors on the results. CCIC is then applied to investigate the short-term dependence characteristics in annual precipitation from 520 stations in China in Section 4, where it is compared with other criteria to illustrate its superiority. Finally, conclusions are given in Section 5.

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2. Correlation Coefficient-Based Information Criterion

2.1. Relationship Between Correlation Coefficient and Serial Dependence

Before quantifying the dependence characteristics of a hydrological time series, it is necessary to first remove deterministic components from the original time series. The abrupt change component can be identified by the Pettitt method or the Brown-Forsythe method (Militino et al., 2020; Ryberg et al., 2020; Wu et al., 2019; Xie, Wu, et al., 2018). The trend component can be detected by the Spearman rank correlation test, Kendall rank correlation test, or linear trend correlation coefficient test (Asfaw et al., 2018; Sang, Sun, et al., 2018; Y.Y. Xie et al., 2016; Yue et al., 2002). The periodic component can be identified by the power spectrum analysis, harmonic analysis, maximum entropy spectrum analysis, or discrete wavelet spectrum (Sang et al., 2021; Xie et al., 2021). For the residual time series x_t ($t = 1, 2, \ldots, p$) which may include dependent components, it can be described by a suitable AR, MA, or ARMA model.

Here, we take the AR model as an example to describe the derivation of the relationship between the correlation coefficient and serial dependence. The residual time series x_t can be described by an autoregressive model AR(p) as:

$$x_t = u + \varphi_1 (x_{t-1} - u) + \varphi_2 (x_{t-2} - u) + \cdots + \varphi_p (x_{t-p} - u) + \varepsilon_t \tag{1}$$

where u is the mean of x_t; $\varphi_1, \varphi_2, \ldots, \varphi_p$ are the autoregressive coefficients; p is the order of the AR model; ε_t is the independent pure random variable with zero mean value and variance σ_ε^2; and ε_t is independent from x_t.

If the dependent component $\varphi_1 (x_{t-1} - u) + \varphi_2 (x_{t-2} - u) + \cdots + \varphi_p (x_{t-p} - u)$ is denoted as η_t, and the random component $u + \varepsilon_t$ is denoted as u_t, then x_t can be expressed as:

$$x_t = \eta_t + u_t \tag{2}$$

where η_t and u_t are independent, and $E(x_t) = u$, $E(u_t) = u$, and $E(\eta_t) = 0$.

The correlation between the time series x_t and the dependent component η_t can be described by the correlation coefficient r:

$$r = \frac{\sum_{t=1}^{n} \left(x_t - \overline{x}\right)\left(\eta_t - \overline{\eta}\right)}{\sqrt{\sum_{t=1}^{n} \left(x_t - \overline{x}\right)^2 \sum_{t=1}^{n} \left(\eta_t - \overline{\eta}\right)^2}} \tag{3}$$

Taking $\overline{x} = u$, $\overline{\eta} = 0$, and $x_t = \eta_t + u_t$, Equation 3 can be simplified as:

$$r = \frac{\sum_{t=1}^{n} {\eta_t}^2 + \sum_{t=1}^{n} (u_t - u)\, \eta_t}{\sqrt{\sum_{t=1}^{n} \left(x_t - \overline{x}\right)^2 \sum_{t=1}^{n} \left(\eta_t - \overline{\eta}\right)^2}} \tag{4}$$

with

$$\sum_{t=1}^{n} (u_t - u)\, \eta_t = nE(u_t \eta_t - u\eta_t) = nE(u_t)\, E(\eta_t) - nuE(\eta_t) = 0 \tag{5}$$

Now r can be described as:

$$r^2 = \frac{\sigma_\eta^2}{\sigma_x^2} \tag{6}$$

where σ_η (σ_x) is the standard deviation of η_t (x_t).

The variance ${\sigma_x}^2$ of x_t is expressed as the sum of the variance ${\sigma_\eta}^2$ of dependent component and the variance ${\sigma_u}^2$ of pure random component:

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$\sigma_x{}^2 = \sigma_\eta{}^2 + \sigma_u{}^2 \tag{7}$$

Combining Equation 6 and Equation 7, one gets:

$$r^2 = 1 - \frac{\sigma_u{}^2}{\sigma_x{}^2} \tag{8}$$

In order to establish the relationship between the correlation coefficient r and the AR model's parameters, the following is further derived.

Multiplying both sides of Equation 1 by $x_t - u$, and taking their expectation (E), one gets:

$$\begin{aligned} E\left(x_t x_t - u x_t\right) = {} & \varphi_1 E\left[\left(x_{t-1} - u\right)\left(x_t - u\right)\right] + \varphi_2 E\left[\left(x_{t-2} - u\right)\left(x_t - u\right)\right] \\ & + \cdots + \varphi_p E\left[\left(x_{t-p} - u\right)\left(x_t - u\right)\right] + E\left(x_t u_t - u u_t\right) \end{aligned} \tag{9}$$

Dividing both sides of Equation 9 by σ_x^2, and considering that

$$\frac{E\left(x_t x_t\right)}{\sigma_x^2} = \frac{D(x) + u^2}{\sigma_x^2}, \ \frac{E\left[\left(x_{t-i} - u\right)\left(x_t - u\right)\right]}{\sigma_x^2} = \rho_i \, (i = 1, 2, \cdots, p) \tag{10}$$

we can obtain:

$$1 = \rho_1 \varphi_1 + \rho_2 \varphi_2 + \cdots + \rho_P \varphi_P + \frac{E\left(u_t x_t\right) - u^2}{\sigma_x^2} \tag{11}$$

as

$$E\left(u_t x_t\right) = E\left(u_t u_t + u_t \eta_t\right) = E\left(u_t u_t\right) = \sigma_u^2 + u^2 \tag{12}$$

Equation 11 can be re-written as:

$$1 = \rho_1 \varphi_1 + \rho_2 \varphi_2 + \cdots + \rho_P \varphi_P + \frac{\sigma_u^2}{\sigma_x^2} \tag{13}$$

Combining Equation 8 and Equation 13, we obtain:

$$r^2 = \rho_1 \varphi_1 + \rho_2 \varphi_2 + \cdots + \rho_P \varphi_P \tag{14}$$

where the autoregressive coefficient $\varphi_i \, (i = 1, 2, \cdots, p)$ can be estimated by the Yule-Walker equations (Li & Jayaweera, 2017):

$$\begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_p \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \cdots & \rho_{p-2} \\ \vdots & \vdots & \cdots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \cdots & 1 \end{pmatrix}^{-1} \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{pmatrix} \tag{15}$$

Therefore, Equation 14 can be expressed by the autocorrelation coefficient $\rho_i (i = 1, 2, \cdots, p)$ as:

$$r^2 = \left(\rho_1 \, \rho_2 \cdots \rho_p\right) \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \cdots & \rho_{p-2} \\ \vdots & \vdots & \cdots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \cdots & 1 \end{pmatrix}^{-1} \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{pmatrix} \tag{16}$$

Since the autocorrelation coefficient $\rho_i (i = 1, 2, \cdots, p)$ is an indicator of the degree of linear dependence in a hydrological time series, it can be used as a preliminary approach to judging whether the time series exhibits

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

dependence characteristics or not. Moreover, Equation 16 shows that the correlation coefficient r between the original time series and its dependent component has already considered the effect of lag-1 to lag-p autocorrelations. Therefore, the correlation coefficient r can be used to judge the dependence of hydrological time series, and further to assess its degree of dependence by comparing it with a threshold r at a certain significance level (such as 5%).

It should be noted that if the MA or ARMA model is used to describe the residual time series x_t, its equation (Equation A1 and (B1), respectively) can be used to substitute Equation 1, and then the relationship between the correlation coefficient r and the model's parameters can also be derived accordingly (shown in Equations A9 and B14, respectively), as shown in the steps in Appendix A and B, respectively. As a result, these derivations indicate that the correlation coefficient r is closely related to serial dependence, and thus it can be a useful indicator to judge the dependence characteristics of hydrological time series as a basis of the CCIC proposed in the following to further determine a suitable model order to describe the dependent components.

2.2. Determination of Suitable Model Order by CCIC

The description of the dependent component in a hydrological time series by regression-based models usually contains two parts: the estimation of parameters, and the determination of the suitable model order. In this study, we focus on the latter. Both AIC and BIC are commonly used for the determination of model order. AIC, also known as the minimum information criterion, was proposed by Akaike (1973) and is an established theory for AR models (Wang et al., 2018). It was further extended to determine the order of ARMA models as well as mixed regression models. AIC is a combination of the maximum likelihood method and information theory, in which the information theory is embodied in the application of Kullback-Leibler (K-L) relative entropy expressed as logarithmic likelihood ratio (Huang et al., 2016). If the order of a regression model is p, the general form of AIC is defined as:

$$\mathrm{AIC}(p) = -2\ln(L(\hat{\beta})) + 2p \tag{17}$$

where $\hat{\beta}$ is the maximum likelihood estimation of parameters, and $\ln(L(\hat{\beta}))$ is the maximum logarithmic likelihood function of the model. If the least square method is used to estimate the residual error function of AIC, then it can be described as (Aho et al., 2014):

$$\mathrm{AIC}(p) = \ln\sigma_\varepsilon^2 + \frac{2p}{n} \tag{18}$$

where σ_ε^2 is the variance of the residual error function, and n is the data length. When $n \to \infty$, the model order determined by AIC cannot converge to the true value according to the probability theory (Fishler et al., 2002). In order to obtain consistent estimation, Akaike (Wang et al., 2018) and Schwarz (Yang et al., 2021) proposed BIC according to the Bayesian principle:

$$\mathrm{BIC}(p) = \ln\sigma_\varepsilon^2 + \frac{\ln n}{n}\mathrm{p} \tag{19}$$

Compared with AIC, the BIC function also includes two parts, where '$\ln n$' replaces the corresponding item '2' in Equation 18. In general, $\ln n >> 2$, thus for a certain time series, the order determined by BIC is often smaller than that determined by AIC.

As explained above, both Equation 18 and Equation 19 include two parts. The former is the variance of the residual error function, reflecting the fitting error of the model; the latter is a penalty term containing the length of time series and the model order, reflecting the uncertainty of the model. Obviously, high accuracy of model fitting is required in practical applications, but higher accuracy means a larger number of parameters which makes the model more complex and results more uncertain. Since both AIC and BIC balance the fitting residual error and uncertainty penalty, it may be logical to choose the optimal model order based on both and evaluate the corresponding parameters at the minimum point of the criterion. However, they have defects and cannot give a reliable determination of model orders in many situations, as discussed earlier.

Sampling errors may exist when calculating the correlation coefficients r in Equation 16 (and also in Equation A9 and (B14)) for limited data samples. In order to judge the reliability of the correlation coefficient calculated, the error of the correlation coefficient is usually estimated statistically (Maesono, 2005). Since the correlation coefficient r can indicate the significance of the dependent component in a time series, its mean square error (σ_r) can be regarded as an indicator of the fitting error of the dependent component, expressed as (Wilson, 2014):

$$\sigma_r = \frac{1 - r^2}{\sqrt{n}} \tag{20}$$

Along with the increase of model order, the mean square error σ_r decreases, implying that the fitting degree between the original time series and its dependent component increases.

Besides, the information entropy index is widely used to quantify the disorder and information of data, which has a positive relationship with the uncertainty of variables (Castillo et al., 2015; Kong et al., 2015; Rajsekhar et al., 2015; Sang, Singh, et al., 2018; Singh, 2013). When investigating uncertainty, greater information entropy of time series indicates larger uncertainty (Koutsoyiannis, 2014; Singh, 2013). Information entropy is defined as:

$$H(X) = H(m_1, m_2, \ldots, m_n) = -c \sum_{i=1}^{n} (m_i \ln m_i) \tag{21}$$

where X denotes any random variable, $H(m_1, m_2, \ldots, m_n)$ is the entropy function, m_i is the probability of the occurrence of the ith information state. For an information system with equal probability, there is $m_1 = m_2 = \ldots = m_n = 1/n$. c is a constant (generally taken as 1), and $\ln(\cdot)$ is the natural logarithm.

Taking AR (1) and AR (2) models as an example again, the following functions are constructed in the form of information entropy, which are related to the model order and can reflect the uncertainty of the model:

$$H(1) = -\left(\frac{1}{n}\ln\frac{1}{n}\right) \tag{22}$$

$$H(2) = \frac{1}{n}\ln\frac{1}{n} - \frac{2}{n}\ln\frac{2}{n} = -\left(\frac{1}{n}\ln\frac{1}{n} + \frac{2}{n}\ln\frac{2}{n}\right) \tag{23}$$

$H(2)$ contains the amount of information in $H(1)$. When $n >> p$, the function indicating the uncertainty of the AR(p) model can be written as:

$$H(p) = -\sum_{k=1}^{p} \left(\frac{k}{n}\right) \ln\left(\frac{k}{n}\right) \tag{24}$$

where p is the model order. For Equation 24, the $H(p)$ value increases along with model order, which means that more parameters will lead to an increase in model uncertainty.

Considering Equation 20 and Equation 24 together, the two components have exactly the same roles as the two terms in AIC and BIC. Therefore, we can use the mean square error of the correlation coefficient to represent the fitting error of the model, and then use the functional form of information entropy as the penalty term. By combining them, we propose a new correlation coefficient-based information criterion (CCIC) for determining the suitable model order:

$$\mathrm{CCIC}(p) = \ln\sigma_r^2 - \sum_{k=1}^{p} \left(\frac{k}{n}\right) \ln\left(\frac{k}{n}\right) \tag{25}$$

For keeping a uniform format with AIC and BIC, the above formula can be further rewritten as:

$$\mathrm{CCIC}(p) = \ln\sigma_r^2 + \frac{1}{n}\sum_{k=1}^{p} k(\ln n - \ln k) \tag{26}$$

Being similar to AIC and BIC, when the CCIC function gets its minimum value, the corresponding order p is taken as the best order for the AR model to be used. Moreover, it should be pointed out that Equation 20 and Equation 26 can also be directly used for the MA and ARMA models, just substituting p by q and $p + q$, respectively. Thus, the proposed CCIC can be applicable to these regression-based models.

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Overall, the specific steps of determining the suitable order by CCIC for regression-based modeling of the dependent components are explained as follows:

1. Identify the deterministic components in the original time series, including abrupt changes, trends, and periodicities, and remove them to obtain the residual time series x_t;
2. Draw the autocorrelation coefficient graph and partial correlation coefficient graph of x_t. If the autocorrelation coefficient graph is tailed (truncated) and the partial correlation coefficient graph is truncated (tailed), then it is a preliminary indication that there are dependent components in x_t, and an AR (MA) model is chosen to describe them; if both the autocorrelation coefficient graph and the partial correlation coefficient graph are tailed, then an ARMA model is chosen to describe the dependent components in x_t;
3. Do the independence test on x_t. If the result shows that the series x_t is random, the model order is zero; otherwise, the existence of dependent components is confirmed, and then the following analysis;
4. Gradually increase the model order p from 1 and estimate the chosen model's parameters by the least square method. The corresponding correlation coefficients r_p between the estimated dependent component η_t and x_t are calculated;
5. For each model order p, calculate the CCIC value using Equation 26. When CCIC obtains the minimum value, the corresponding correlation coefficient is denoted as r^*, and the corresponding model order is denoted as p^*;
6. Take p^* as the estimated model order, and use the chosen model to describe the time series x_t, and take the difference between x_t and the modeling results as the residual error;
7. Do the same analysis of the residual error following step (3). If the residual error still includes a dependent component, set $p^* = p^* + 1$ to re-extract the residual error following step (6) and update r^* accordingly, until the residual error shows independent characteristics;
8. Finally, the reasonability of the most suitable model order p^* is confirmed, and the significance of the dependent component is confirmed by comparing r^* with the threshold r at the chosen significance level (5% level is used in this study).

Besides, Figure 1 also shows the flowchart for determining the suitable model order by the proposed CCIC.

3. Verification of CCIC by Monte-Carlo Experiments

3.1. Rationality of Correlation Coefficient as an Index for Quantifying Dependence

The above derivations illustrate that the correlation coefficient r between the original time series and its dependent components can be used to quantify the dependence of hydrological time series. Here, we design the Monte-Carlo experiments by taking the first-order, second-order, and third-order AR models as examples, to validate the reliability of the derived relationship between correlation coefficient r and serial dependence.

For the first-order AR model (i.e., AR(1) model), its expression can be written as $x_t = \varphi_1 (x_{t-1} - u) + u_t$. Based on the Yule-Walker equations, we obtain $\rho_1 = \varphi_1$ and insert into Equation 14, then r^2 can be shown as:

$$r^2 = \varphi_1^2 = \rho_1^2 \tag{27}$$

We use the Monte-Carlo method to generate synthetic time series that follows the Pearson-III probability distribution, which is widely considered for hydrological frequency analysis. Each time series has the same length n of 1,000, initial value $x_1 = 100$, mean value $u = 100$, variation coefficient $C_{uv} = 0.2$, and skewness coefficient $C_{us} = 0.4$. The specific steps are described below:

1. Generate the pure random series u_t that satisfy the above conditions and generate the dependent series η_t by setting its autoregressive coefficients φ_1 as certain values. Here, we set 18 situations as $\varphi_1 = \pm 0.1, \pm 0.2, ..., \pm 0.9$, respectively;
2. Get the synthetic time series x_t by combining η_t and φ_1 for each situation;
3. Use Equation 3 to calculate the correlation coefficient x_b ($b = 1,2,\ldots 18$) between x_t and η_t after removing the random series u_t from x_t. Each case is repeated 1,000 times (i.e., $i = 1,2,\ldots,1{,}000$) to ensure the stability of results, and the average value of r_b in each group of experiments is calculated as:

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

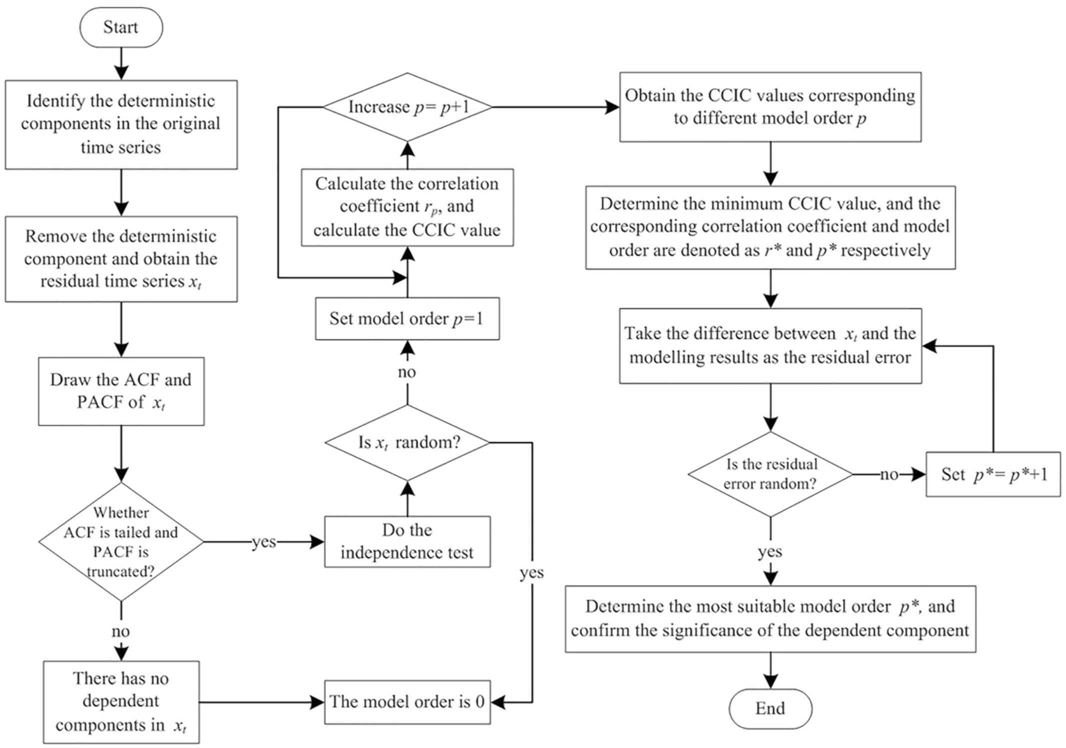

Figure 1. Flowchart of determining the suitable model order by the proposed correlation coefficient-based information criterion (CCIC), for regression-based modeling of the dependent components in hydrological time series.

$$r_b = \frac{1}{1000}\sum_{i=1}^{1000} r_{ib}(b = 1, 2, ...18) \tag{28}$$

4. Under the stability condition (Nagakura, 2009) of AR(1) model with $|\varphi_1| < 1$, the correlation coefficient can be directly calculated by Equation 27, which is denoted as r_a (a = 1,2,…18);
5. The relative error between r_a and r_b is calculated as: $\delta = \left(\frac{|r_b - r_a|}{r_a}\right) \times 100\,(\%)$.

The AR(2) model is expressed as: $x_t = \varphi_1 (x_{t-1} - u) + \varphi_2 (x_{t-2} - u) + u_t$, with $\rho_1 = \frac{\varphi_1}{1-\varphi_2}$ and $\rho_2 = \frac{\varphi_1^2}{1-\varphi_2} + \varphi_2$ from the Yule-Walker equations in Equation 14, we can obtain Equation 29:

$$r^2 = \frac{\varphi_1^2 (1 + \varphi_2)}{1 - \varphi_2} + \varphi_2^2 = \frac{(1 - 2\rho_2)\rho_1^2 + \rho_2^2}{1 - \rho_1^2} \tag{29}$$

Similarly, for the AR(3) model, we can also obtain the following equation:

$$\begin{aligned} r^2 &= \frac{\varphi_1 (\varphi_1 + \varphi_2\varphi_3)}{1 - \varphi_2 - \varphi_1\varphi_3 - \varphi_3^2} + \frac{\varphi_2 \left(\varphi_1^2 + \varphi_1\varphi_3 - \varphi_2^2 + \varphi_2\right)}{1 - \varphi_2 - \varphi_1\varphi_3 - \varphi_3^2} \\ &\quad + \frac{\varphi_3 \left(\varphi_1^3 + \varphi_1^2\varphi_3 - \varphi_1\varphi_2^2 + 2\varphi_1\varphi_2 - \varphi_1\varphi_3^2 + \varphi_2^2\varphi_3 - \varphi_2\varphi_3 - \varphi_3^3 + \varphi_3\right)}{1 - \varphi_2 - \varphi_1\varphi_3 - \varphi_3^2} \\ &= \frac{-\rho_1 \left(-\rho_1^3 + \rho_3\rho_1^2 + \rho_1\rho_2^2 - \rho_1\rho_2 + \rho_1 - \rho_2\rho_3\right)}{(\rho_2 - 1)\left(-2\rho_1^2 + \rho_2 + 1\right)} + \frac{\rho_2 \left(-\rho_1^2 - \rho_1\rho_3 + \rho_2^2 + \rho_2\right)}{-2\rho_1^2 + \rho_2 + 1} \\ &\quad + \frac{\rho_3 \left(\rho_1^3 - \rho_3\rho_1^2 + \rho_1\rho_2^2 - 2\rho_1\rho_2 + \rho_3\right)}{2\rho_1^2\rho_2 - 2\rho_1^2 - \rho_2^2 + 1} \end{aligned} \tag{30}$$

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

10.1029/2021WR031606

Table 1
r_a, r_b, and Their Relative Error δ Under Different Parameters in AR(1) Model

φ_1	r_a	r_b	δ	φ_1	r_a	r_b	δ	φ_1	r_a	r_b	δ
−0.9	0.9	0.898	0.23%	−0.3	0.3	0.300	0.03%	0.4	0.4	0.399	0.34%
−0.8	0.8	0.798	0.20%	−0.2	0.2	0.202	0.80%	0.5	0.5	0.499	0.22%
−0.7	0.7	0.699	0.17%	−0.1	0.1	0.100	0.25%	0.6	0.6	0.597	0.47%
−0.6	0.6	0.599	0.10%	0.1	0.1	0.099	0.87%	0.7	0.7	0.696	0.53%
−0.5	0.5	0.499	0.29%	0.2	0.2	0.199	0.57%	0.8	0.8	0.796	0.45%
−0.4	0.4	0.399	0.21%	0.3	0.3	0.297	1.13%	0.9	0.9	0.896	0.46%

In the Monte-Carlo experiments, the parameters of AR models are set to satisfy the stability conditions, whose reasonable ranges were determined by the eigenvalue method in this study (Slawski & Hein, 2015). We randomly selected 18 associative groups of autoregressive coefficients, as explained above, and the relative errors under different situations were calculated based on the above steps. Results are given in Tables 1–3 show that the relative errors between r_a and r_b are all within 1.25% for all the three AR model situations. It indicates that Equation 14 is reliable and the correlation coefficient r can be used as an effective index to describe the dependence characteristics of time series.

3.2. Verification of the Efficiency of CCIC

Here, we used the same synthetic time series generated in Section 3.1 to compare the accuracy of CCIC with that of AIC and BIC and illustrate the efficacy of CCIC. Taking AR(1), AR(2), and AR(3) models as examples again, the following steps were used for the experiments:

1. Randomly generate the autoregressive coefficients φ_1 of the dependent time series under stability conditions, and set the data length n as 50, 75, 100, 150, and 200, respectively;
2. When $n = 50$, x_t is generated by combining the dependent time series with the random time series u_t following the assumptions in Section 3.1, with the initial value $x_1 = 100$;
3. Predefine the model order range from 1 to $n/3 + 1$ (Jirak, 2012). We can identify the minimum value of the three criteria (CCIC, AIC, BIC) to select the corresponding optimal model orders based on the residual variance σ_ε^2 and the correlation coefficient r between the AR-simulated time series and the original time series.
4. To ensure the reliability of results, 1,000 synthetic time series with different values of φ_1 are randomly generated. Follow steps (2) and (3) to count the number of models whose orders are evaluated as 1. By dividing the number by 1,000, we can obtain the accuracy of model order by the three criteria, respectively.
5. Repeat the above steps and change the value of n to assess the accuracy of the three criteria. Take its average value as the accuracy of the first-order AR model from three criteria.
6. Do the same analysis above for the AR(2) and AR(3) models to obtain the accuracy of results from the three criteria.

Table 2
r_a, r_b, and Their Relative Error δ Under Different Parameters in AR(2) Model

φ_1	φ_2	r_a	r_b	δ	φ_1	φ_2	r_a	r_b	δ
0.1	−0.8	0.801	0.798	0.38%	−0.3	−0.7	0.711	0.710	0.24%
−0.7	0.2	0.880	0.878	0.29%	−0.7	−0.5	0.643	0.642	0.08%
0.3	−0.4	0.446	0.446	0.07%	0.6	−0.8	0.825	0.822	0.29%
0.8	0.1	0.890	0.885	0.59%	−0.1	−0.5	0.503	0.503	0.08%
0.6	−0.8	0.825	0.822	0.27%	0.2	−0.4	0.421	0.419	0.40%
0.6	−0.9	0.911	0.908	0.25%	−0.2	0.7	0.847	0.840	0.77%
0.5	−0.5	0.577	0.576	0.16%	−0.3	−0.1	0.289	0.292	0.79%
−0.2	−0.1	0.207	0.208	0.62%	0.5	0.1	0.562	0.558	0.59%
0.5	0.3	0.745	0.738	0.94%	−0.5	0.2	0.644	0.643	0.27%

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3
r_a, r_b, and Their Relative Error δ Under Different Parameters in AR(3) Model

φ_1	φ_2	φ_3	r_a	r_b	δ	φ_1	φ_2	φ_3	r_a	r_b	δ
0.1	0.7	−0.2	0.729	0.726	0.36%	−0.4	0.7	0.6	0.833	0.823	1.25%
0.6	−0.5	0.7	0.769	0.764	0.67%	0.3	−0.7	0.6	0.886	0.882	0.41%
−0.9	−0.1	0.4	0.848	0.847	0.09%	0.7	−0.9	0.4	0.799	0.797	0.27%
0.3	−0.6	−0.3	0.805	0.802	0.37%	0.1	0.1	0.4	0.465	0.461	0.89%
0.2	0.1	−0.1	0.244	0.241	0.97%	−0.1	−0.6	−0.1	0.602	0.603	0.14%
0.6	−0.6	0.2	0.600	0.599	0.15%	0.5	0.5	−0.4	0.702	0.700	0.35%
−0.2	0.3	−0.4	0.809	0.802	0.97%	−0.3	0.5	0.1	0.641	0.639	0.32%
0.3	0.5	−0.7	0.820	0.817	0.44%	0.3	−0.8	−0.3	0.982	0.978	0.36%
−0.2	−0.5	−0.6	0.768	0.765	0.39%	−0.3	−0.7	0.1	0.754	0.753	0.24%

Table 4 shows the accuracy of correctly identifying the true order of AR models. For the AR(1) model, the average accuracy of CCIC was higher than that of BIC and much higher than that of AIC. When n changed from 50 to 200, the accuracy of CCIC was above 90%. For the AR(2) model, CCIC had the highest accuracy (80.06%) among the three criteria, especially for long data lengths. When CCIC was applied to the AR(3) model, its average accuracy was 63.56%, still keeping higher accuracy than that of AIC and BIC for different data lengths. Hence, it was found that CCIC had high efficiency for the determination of suitable AR(p) model order, especially for AR(1) and AR(2) models.

In order to further verify the applicability of the proposed CCIC, we also did Monte-Carlo experiments for the MA and ARMA models. Details can be found in Appendixes A and B, respectively. The results in Tables A1 and B1 also indicated that the CCIC had higher efficiency for the determination of suitable model order compared to AIC and BIC, no matter which data lengths are considered, being consistent with the results for AR models. It is just due to the same regression-based essence of these models. Overall, all the results verified higher and more stable accuracy of determining the true model order by CCIC than by AIC and BIC, and thus the proposed CCIC can be applicable for the AR, MA, and ARMA models.

3.3. Influence of Main Factors on the Verification Test

Different parameters are involved in the calculation of CCIC, which may impact the accuracy of quantification of dependence characteristics. To investigate how the four main factors (mean value u, variation coefficient C_{uv}, skewness coefficient C_{us}, and initial value x_1) influenced the calculation of the correlation coefficient r and the accuracy of CCIC, we designed a set of Monte-Carlo (MC) experiments, by taking AR(1) model as an example again. For example, the values of u, C_{uv} and C_{us} were preset and fixed, and then x_1 varied to investigate their influence on r. We repeated 1,000 MC experiments for each case of the four factors, and take the mean value as the final value of r and the accuracy of CCIC.

Table 4
Accuracy of the Determination of AR Model Orders Under Different Data Lengths by AIC, BIC, and the Proposed CCIC

	AR(1)			AR(2)			AR(3)		
Data length	AIC	BIC	CCIC	AIC	BIC	CCIC	AIC	BIC	CCIC
$n = 50$	30.18%	81.99%	90.36%	22.34%	61.58%	74.49%	19.03%	46.80%	54.03%
$n = 75$	31.18%	88.82%	93.22%	24.51%	69.37%	77.33%	21.03%	57.77%	59.50%
$n = 100$	34.04%	90.81%	94.16%	28.60%	74.79%	80.44%	23.57%	61.47%	63.47%
$n = 150$	35.65%	93.22%	95.29%	32.17%	78.54%	83.23%	28.00%	67.87%	69.13%
$n = 200$	40.31%	94.67%	96.75%	37.05%	80.46%	84.79%	30.63%	70.50%	71.67%
Mean value	34.27%	89.90%	93.96%	28.93%	72.95%	80.06%	24.45%	60.88%	63.56%

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

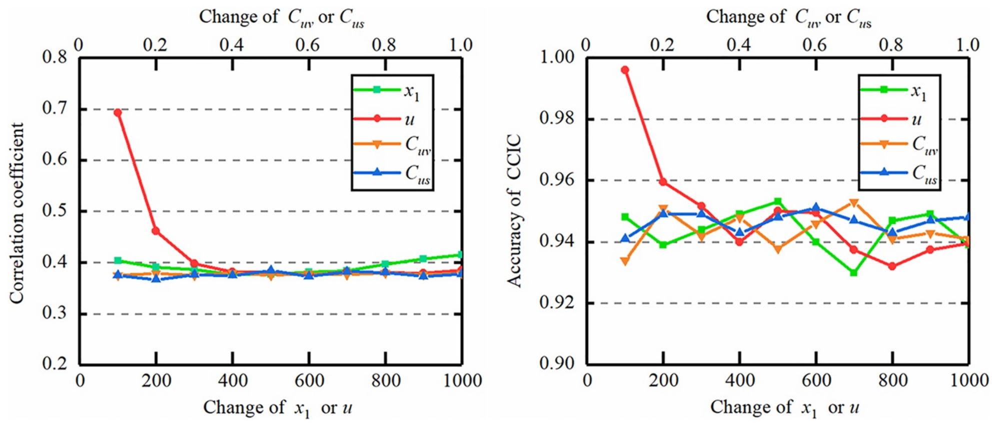

Figure 2. Changes of correlation coefficient (left) and accuracy of correlation coefficient-based information criterion (right) with the four parameters of mean value u, variation coefficient C_{uv}, skewness coefficient C_{us}, and initial value x_1.

Results in Figure 2 (left) show that the initial value of x_1 and the statistical parameters (C_{uv} and C_{us}) of pure random components had little impact on r. However, with the increase of u, r showed first a significant downward trend and then remained stable after u was close to x_1. Figure 2 (right) illustrates that the influence of the other three parameters except u was weak on the accuracy of CCIC; when the value of u increased, the accuracy of CCIC showed a downward trend but within a 7% variation range.

We further analyzed the factors that influenced the quantification of dependent components. Combining Equation 6 and Equation 7, we can get:

$$r^2 = \frac{1}{1 + \frac{\sigma_u^2}{\sigma_\eta^2}} \tag{31}$$

where σ_u^2 represents the variance of the pure random component and σ_η^2 represents the variance of the dependent components. Since the signal-to-noise ratio (SNR) represents the power spectrum ratio of signal to noise, the formula for estimating the SNR of a time series with dependent components can be written as (Herrick, 2014; Xie, Zhao, et al., 2018):

$$SNR = \frac{\sigma_\eta^2}{\sigma_u^2} \tag{32}$$

Thus, according to Equation 31 and Equation 32, the relationship between r and SNR can be expressed as:

$$r^2 = \frac{1}{1 + \frac{1}{SNR}} \tag{33}$$

Equation 33 shows that SNR of a time series has a positive correlation with the absolute value of the correlation coefficient. Moreover, as $\sigma_u^2 = u^2 C_{uv}^2$, it can be deduced that u has a negative correlation with r, which is consistent with the results shown in Figure 2**(left).** To visually show the relationship between SNR and r, as well as the relationship between the accuracy of CCIC and SNR, we randomly generated 1,000 sets of first-order autoregressive time series to obtain the results, as shown in Figure 3, which verified the reasonability of Equation 33 (blue curve). Moreover, it shows that there were little changes in the accuracy of CCIC along with the increase of SNR (green curve), indicating the little effect of SNR on the determination of model order by CCIC.

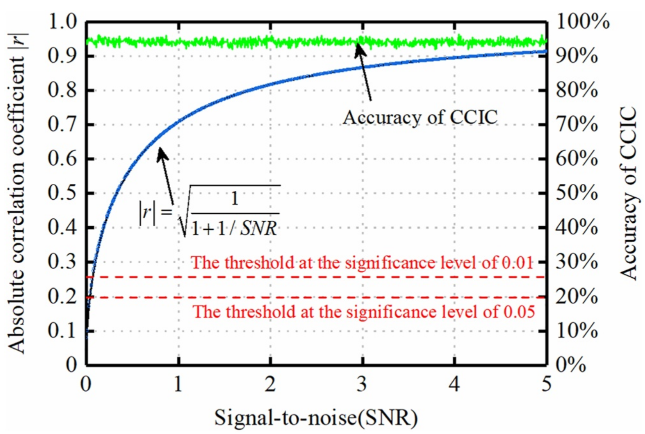

Figure 3. Scatter diagram showing the relationship between SNR and the absolute value of r, and the relationship between the accuracy of correlation coefficient-based information criterion and signal-to-noise ratio. Here 1,000 sets of first-order autoregressive time series were used. The red dotted lines indicate the thresholds of the correlation coefficient at 5% and 1% significance levels.

10.1029/2021WR031606

Considering the same regression-based essence of these models and their consistent results in Table 4, Table A1, and Table B1, the influences of these factors on the MA and ARMA models are not repeated here. To sum up, the main factor impacting the identification of dependent components is the magnitude of the mean value of the time series to be analyzed. If the mean value of the time series is close to its initial value, the calculation of r and the accuracy of CCIC is weakly affected by the four main factors mentioned (u, C_{uv}, C_{us}, x_1). Besides, a higher SNR value suggests a larger proportion of dependent components, which is easier to be identified. However, the signal-to-noise ratio will cause little difference in the determination of model order by CCIC.

4. Detection of Short-Term Dependence Characteristics in Annual Precipitation in China

To further verify the efficacy of the proposed CCIC, the annual precipitation data observed at 520 meteorological stations over China from 1961 to 2013 were collected for investigating the short-term dependence characteristics of annual precipitation. The data were obtained from the China Meteorological Data Sharing Service System (http://cdc.cma.gov.cn/). Considering the tailed characteristics of autocorrelation coefficient graphs of the precipitation time series (with examples shown in following Figure 7), we used the AR model for describing the dependent components in them and applied the proposed CCIC to determine the optimal model order for each time series. According to the method developed in Section 2.2, the deterministic components of the precipitation time series were first checked and removed, and the residual precipitation time series was obtained. After that, the maximum AR model order was set as $n/3 + 1$ ($n = 53$ here). By estimating the values of the AR model's parameters, the correlation coefficients between the residual precipitation time series and its dependent components were calculated, and the CCIC values under different model orders were also calculated. We finally selected the model order corresponding to the minimum CCIC value as the optimal model order.

Figure 4 visually shows the spatial difference in the degree of dependence of annual precipitation at the 520 stations. According to the results, annual precipitation at only 38 stations (7.3% of the total 520 stations) contained

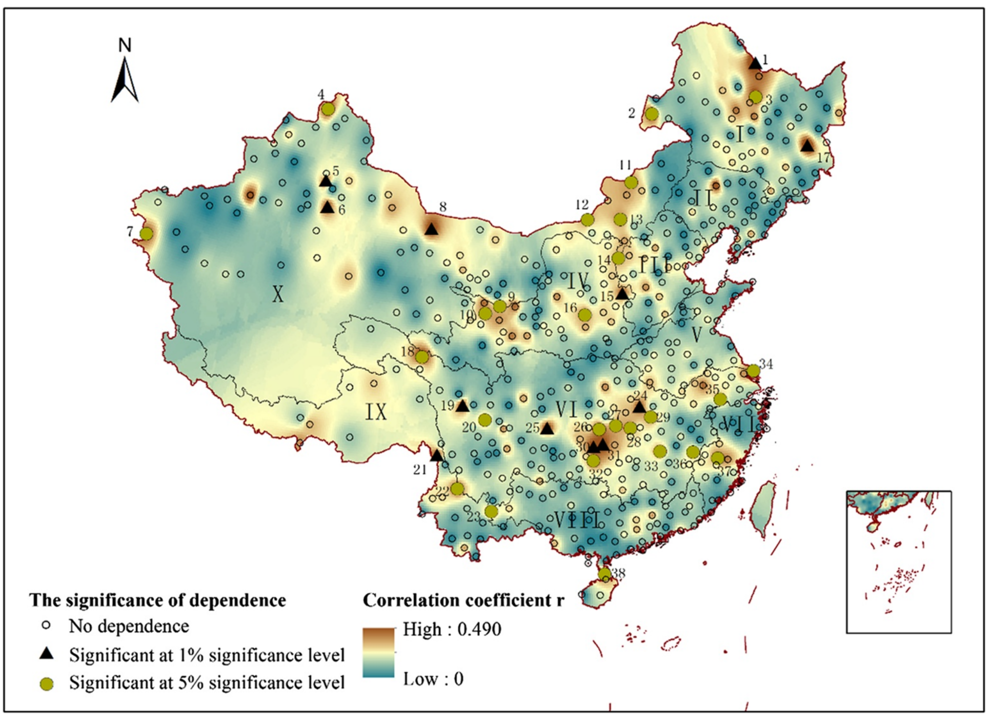

Figure 4. Significance of dependence characteristics for the annual precipitation time series at 520 stations in China. I, the Songhua River Basin; II, the Liaohe River Basin; III, the Haihe River Basin; IV, the Yellow River Basin; V, the Huaihe River Basin; VI, the Yangtze River Basin; VII, the Southeast River Basin; VIII, the Pearl River Basin; IX, the Southwest River Basin; and X, the Northwest River Basin.

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

10.1029/2021WR031606

the dependence components at the 5% significance level. These annual precipitation time series with significant short-term dependence characteristics were mainly located in the Songhua River basin, Northwest China, the Yellow River basin, and especially the whole mid and lower reaches of the Yangtze River basin. It may be due to the combined influence of geographic and topological conditions, and the complex Asian monsoon effects (Jiang et al., 2017; Markonis & Koutsoyiannis, 2016; Sang, Singh, et al., 2018), while more detailed physical causes should be further explored.

Figure 5 shows the determination of AR model orders for the 38 annual precipitation time series that contain dependent components. It indicates that the results of CCIC were consistent with those by BIC in most cases, but AIC gave much higher model orders. By using the AR model orders determined by CCIC, the residual errors of these precipitation time series were obtained, and their correlation coefficients were not significant at the 5% significant level and thus showed independent characteristics, implying the reasonability for the description of dependent components by the AR models determined. Moreover, Figure 6 shows the correlation coefficients of dependent components determined by the three criteria (AIC, BIC, and CCIC) and the thresholds at two significance levels, where $r_{5\%}$ was the threshold at the significance level of 5% and $r_{1\%}$ was the threshold at the significance level of 1%. It clearly indicated that the dependent components determined by AIC were not significant at many stations and that BIC was also not significant in several stations. However, the dependent components determined by CCIC were significant at all 38 stations, which further proved the rationality of the proposed CCIC.

In order to verify the reasonability of the CCIC results, we randomly selected four stations from the 38 stations as examples for further explanations, which were Jishou (JS), Youyu (YY), Mudanjiang (MD), and Qingshui River (QS) stations, with the station number of 30, 14, 17, and 18, respectively (as shown in Figure 4). The autocorrelation coefficient graphs and partial correlation coefficient graphs of the residual precipitation time

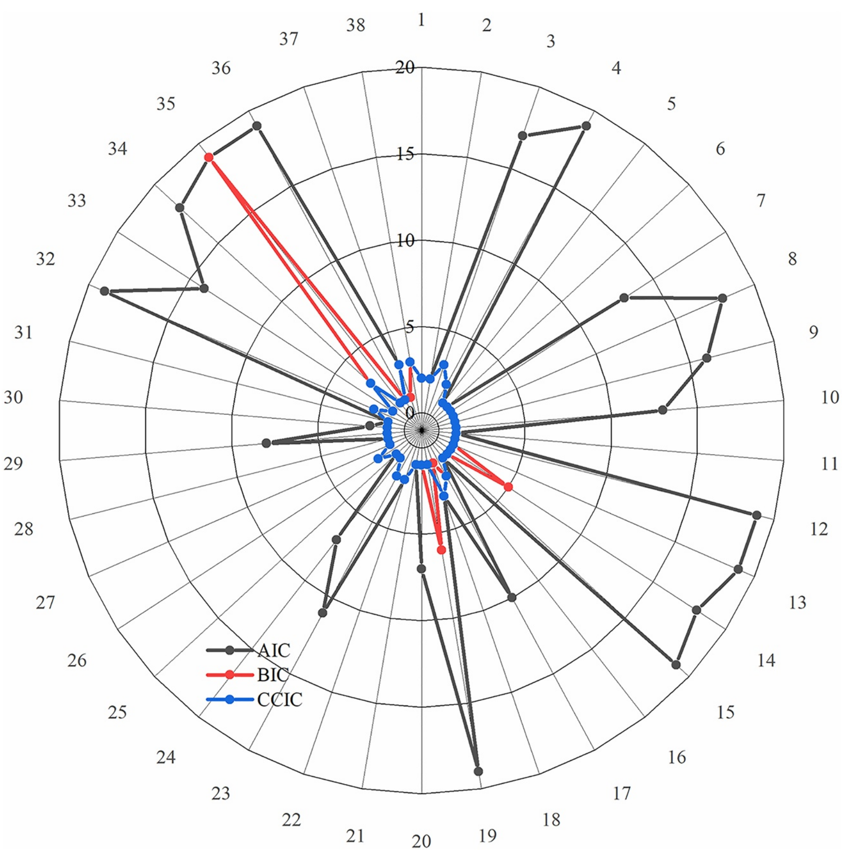

Figure 5. Selection of auto-regressive model orders by the three criteria (Akaike Information Criterion, Bayesian Information Criterion, and correlation coefficient-based information criterion) for the annual precipitation time series that contains dependent components at 38 stations in China.

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

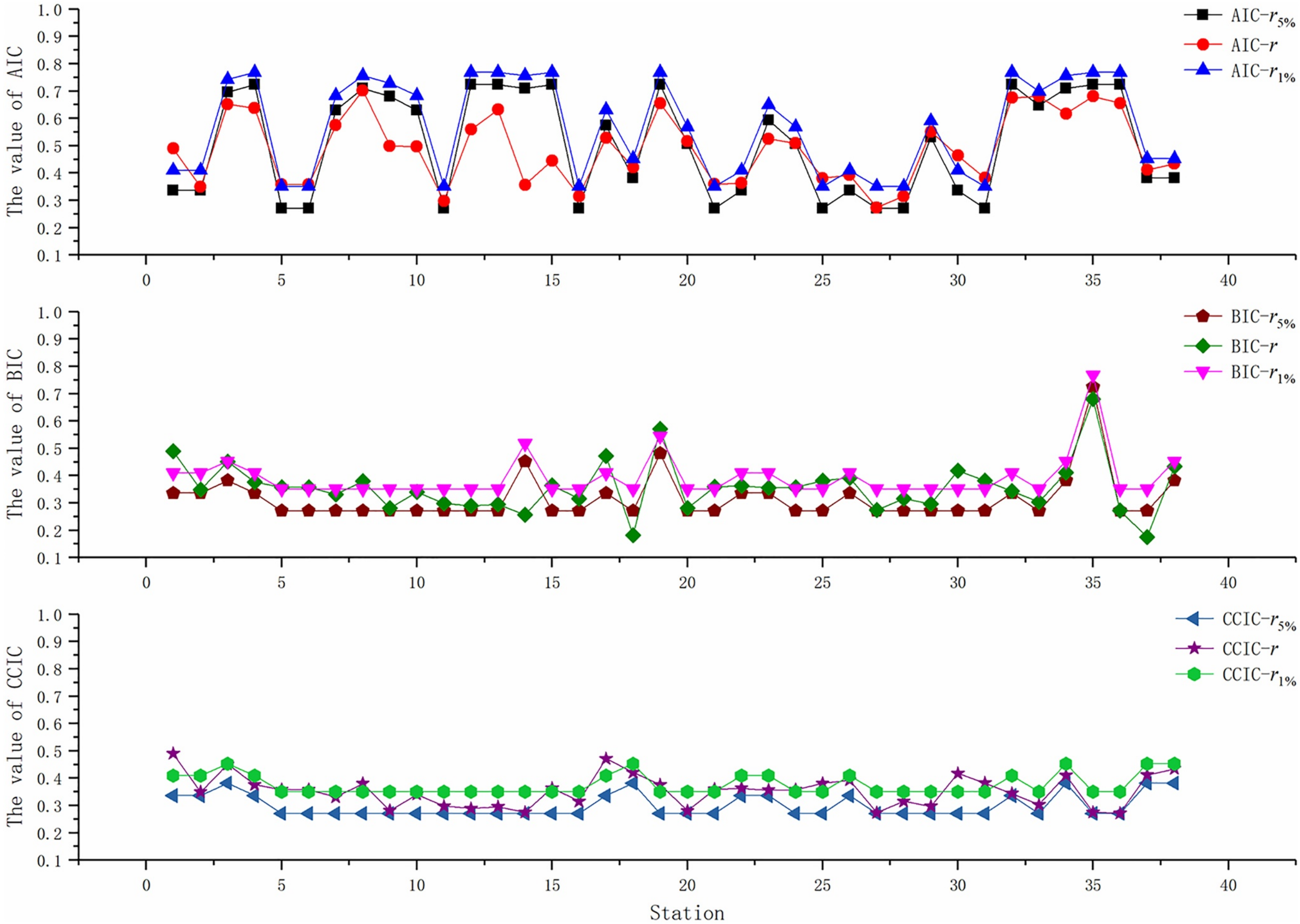

Figure 6. Correlation coefficients (r) of the dependent components in the annual precipitation time series at 38 stations determined by different criteria (Akaike Information Criterion, Bayesian Information Criterion, and correlation coefficient-based information criterion).

series were plotted, as shown in Figure 7. The red line in the figure is the upper and lower thresholds at a 5% significance level. If the autocorrelation coefficient was within the permissible limits, the time series was considered to be independent, otherwise, the time series had dependent components. Figure 7 indicates that except for the YY Station, there were dependent components in the other three annual precipitation time series. It can be considered that the annual precipitation time series at JS, YY, MD, and QS stations were truncated with lag-1, lag-1, lag-2, and lag-3 partial correlation coefficients, respectively. This was consistent with the results of CCIC. However, AIC results were 2, 18, 10, 3, and BIC ranking results were 1, 5, 2, and 1, respectively. Thereby, it was deducted that CCIC performed better than the other two criteria for determining the suitable model orders.

The following principle was used to further verify the rationality of the model order determined by the three criteria. Specifically, the suitable model order should make sure that the time series had dependence characteristics but its residual error was independent. The changes of correlation coefficient r of each precipitation time series along with model order increase are plotted in Figure 8, where the upper and lower red lines represented the thresholds of r at 1% and 5% significance levels, respectively. If the correlation coefficients reached beyond the dotted red line, it was thought that the time series for this model order had dependence characteristics, and vice versa. As shown in Figure 8 (a), the annual precipitation time series at JS station in the first eight order had dependence characteristics. When the model order was one, the order was minimum and the model was the simplest, with its residual error indicating independent characteristics, so it satisfied the principle explained above. In Figure 8 (b), the annual precipitation time series at the YY station had dependence characteristics only

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

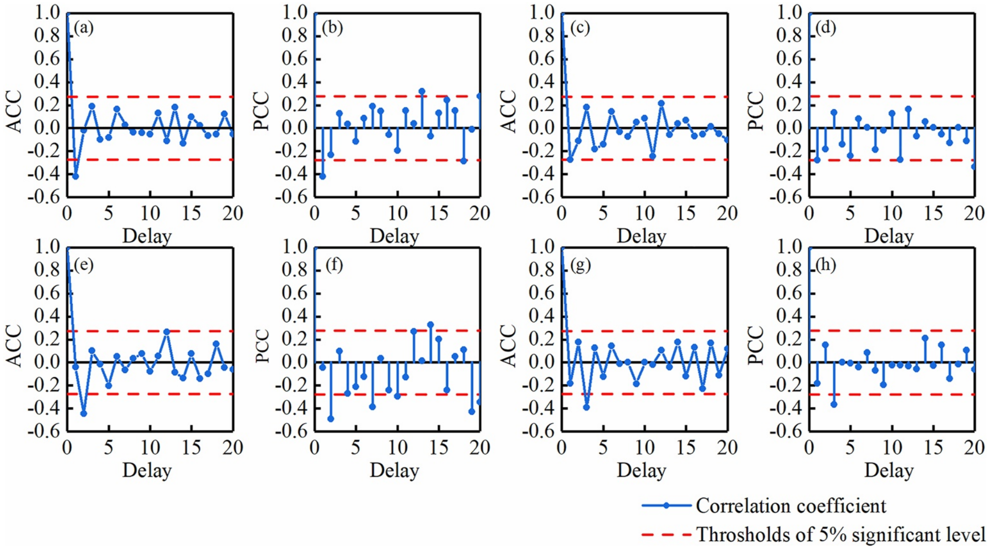

Figure 7. Autocorrelation coefficient (denoted as ACC) graphs and partial correlation coefficient graphs (denoted as PCC) of the selected four precipitation time series. (a), (c), (e) and (g) are ACC graphs of JS, YY, MD, and QS station, respectively. (b), (d), (f) and (h) are PCC graphs of JS, YY, MD, and QS station, respectively.

in the first order, and its residual error was independent, which met the principle. Similarly, as shown in Figure 8 (c) and (d), when the precipitation time series at MD and QS stations had orders 2 and 3, respectively, the minimum orders that satisfied the principle were obtained. It proved the rationality of the model order determination by CCIC, differing from the results of AIC and BIC.

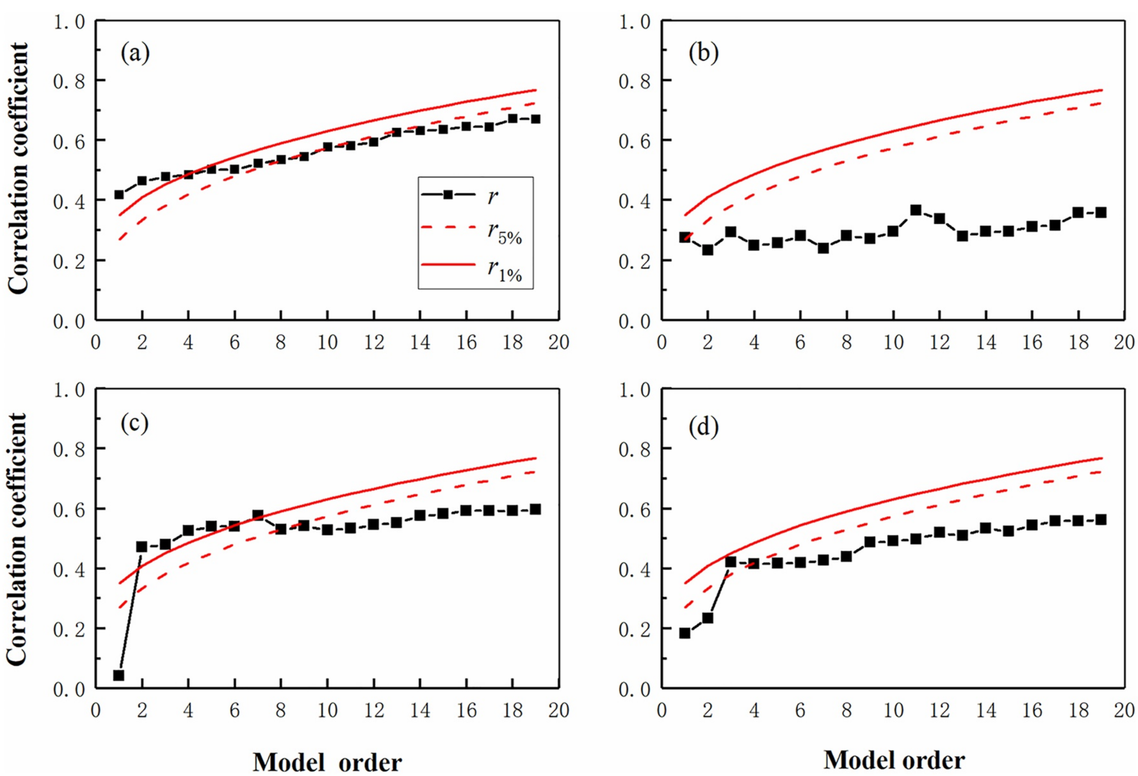

Figure 8. Changes of correlation coefficient © of the four precipitation time series with the increase of model order. (a) JS station, (b) YY station, (c) MD station, and (d) QS station.

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

5. Conclusion

In this study, a method called CCIC was proposed by combining the correlation coefficient r and information theory to determine a suitable model order for describing the dependent components in hydrological time series using regression-based models. Results verified the superiority of CCIC compared to AIC and BIC which had been used widely. According to the results of different Monte-Carlo experiments, the rationality of correlation coefficient r, as the key index in CCIC, for quantifying the significance of dependence had been verified. Moreover, results showed that the estimation of the correlation coefficient was mainly influenced by both the magnitude of mean value and the signal-to-noise ratio, and bigger correlation coefficient r responses to a higher signal-to-noise ratio of time series, implying more significance of dependent components. However, only the magnitude of the mean value had big impact on the accuracy of the CCIC estimation. If the mean value of the time series was close to its initial value, the estimation of CCIC (and also r) was weakly influenced by the four main factors of the mean value (u), variation coefficient (C_{uv}), skewness coefficient (C_{us}), and initial value (x_1), which led to high accuracy of the determination of suitable model order, especially for long data length.

By applying the proposed CCIC to analyze the annual precipitation at 520 meteorological stations in China, its advantage compared to AIC and BIC was further verified. High efficiency for the determination of model order can be obtained by CCIC rather than AIC and BIC. Moreover, CCIC can also more effectively identify the dependent components in hydrological time series by following the principle that "the suitable model order should make sure that the time series has dependence characteristics but its residual error is independent." However, the dependent components determined by AIC and BIC cannot be guaranteed to be significant in all situations, violating the above principle. Besides, it was found that the significant short-term dependence characteristics in annual precipitation time series mainly occurred in local regions especially in the Yangtze River basin, while the physical causes should be further explored.

Overall, the advantages of the proposed CCIC for the determination of suitable model order rather than AIC and BIC were confirmed in this study. It should be pointed out that based on the main idea of the proposed CCIC, it could have a wide application range, just like the wide applications of AIC and BIC. The application of the CCIC approach in other hydrologic cases (runoff, flood peak, *etc.*) and other research areas, as well as different scales, is worth further exploring, to further confirm its reliability and superiority.

Appendix A: Results of Monte-Carlo Experiments for MA Models

Using a moving average model MA(q), the residual time series x_t can be described as:

$$x_t = u + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q} \tag{A1}$$

where u is the mean of x_t; $\theta_1, \theta_2, \ldots, \theta_q$ are the moving average coefficients; q is the order of the MA model; ε_t has the same meaning as that in Equation 1.

If the dependent component $-\theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q}$ is denoted as η_t, and the random component $u + \varepsilon_t$ is denoted as u_t, then x_t can be expressed as:

$$x_t = \eta_t + u_t \tag{A2}$$

where η_t and u_t are independent, and $E(x_t) = u$, $E(u_t) = u$, and $E(\eta_t) = 0$.

Then, the same derivations as that in Equations 2–8 can also be obtained accordingly to get the following equation:

$$r^2 = 1 - \frac{{\sigma_u}^2}{{\sigma_x}^2} \tag{A3}$$

In order to establish the relationship between the correlation coefficient r and the MA model's parameters, the following is derived.

Squaring both sides of Equation A1, and taking their expectation (E), one gets:

$$E(x_t x_t) = E\left[(u_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q})(u_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q})\right] \tag{A4}$$

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Considering that

$$E\left(x_t x_t\right) = D(x) + u^2 \tag{A5}$$

and

$$\begin{aligned} E\left[\left(u_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q}\right)^2\right] &= D\left(u_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q}\right) \\ &+ \left[E\left(u_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q}\right)\right]^2 \end{aligned} \tag{A6}$$

Equation A4 can be re-written as:

$$\sigma_x^2 + u^2 = \left(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\right)\sigma_\varepsilon^2 + u^2 \tag{A7}$$

After simplification, we can get:

$$\sigma_x^2 = \left(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\right)\sigma_\varepsilon^2 \tag{A8}$$

As $\sigma_\varepsilon^2 = \sigma_u{}^2$, we combine Equation A3 and Equation A8 to obtain:

$$r^2 = 1 - \frac{1}{\left(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\right)} \tag{A9}$$

The above equation clearly indicates the relationship between the correlation coefficient r and the parameters of the MA model.

We design the Monte-Carlo experiments by taking the first-order, second-order, and third-order MA models as examples to compare the accuracy of CCIC with that of AIC and BIC. The synthetic time series has initial value $x_1 = 100$, mean value $u = 100$, variation coefficient $C_{uv} = 0.2$, and skewness coefficient $C_{us} = 0.4$, and the data length n is set as 50, 75, 100, 150, and 200, respectively. For the other parameter values and the experiment steps, they are just the same as that for the AR models.

For the first-order MA model (i.e., MA(1) model), its expression can be written as $x_t = -\theta_1\varepsilon_{t-1} + u_t$. Substituting the first-order autocorrelation coefficient formula $\rho_1 = \frac{-\theta_1}{1+\theta_1^2}$ into Equation A9, then r^2 can be shown as:

$$r^2 = 1 + \frac{\rho_1}{\theta_1} \tag{A10}$$

The MA(2) model is expressed as: $x_t = -\theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} + u_t$. Substituting the second-order autocorrelation coefficient formula $\rho_2 = \frac{-\theta_2}{1+\theta_1^2+\theta_2^2}$ into Equation A9, r^2 can be shown as:

$$r^2 = 1 - \frac{1}{1 + \theta_1^2 + \theta_2^2} = 1 + \frac{\rho_2}{\theta_2} \tag{A11}$$

Similarly, for the MA(3) model, we can also obtain the following equation:

Table A1
Accuracy of the Determination of MA Model Orders Under Different Data Lengths by AIC, BIC, and the Proposed CCIC

	MA(1)			MA(2)			MA(3)		
Data length	AIC	BIC	CCIC	AIC	BIC	CCIC	AIC	BIC	CCIC
$n = 50$	57.33%	70.67%	94.67%	44.40%	53.93%	61.53%	35.20%	39.00%	44.33%
$n = 75$	54.27%	73.93%	95.13%	44.07%	55.93%	65.67%	34.13%	43.27%	51.07%
$n = 100$	52.53%	75.80%	96.33%	46.00%	60.47%	67.60%	32.27%	42.80%	53.60%
$n = 150$	51.67%	77.20%	97.13%	43.07%	61.20%	72.33%	32.67%	45.00%	54.73%
$n = 200$	50.27%	78.60%	97.33%	45.47%	63.47%	74.53%	30.60%	43.87%	58.53%
Mean value	53.21%	75.24%	96.12%	44.60%	59.00%	68.33%	32.97%	42.79%	52.45%

$$r^2 = 1 - \frac{1}{1 + \theta_1^2 + \theta_2^2 + \theta_3^2} = 1 + \frac{\rho_3}{\theta_3} \tag{A12}$$

Table A1 shows the accuracy of correctly identifying the true order of MA models. For the MA(1) model, the average accuracy of CCIC was higher than that of BIC and much higher than that of AIC. When n changed from 50 to 200, the accuracy of CCIC was above 90%. For the MA(2) model, CCIC had the highest accuracy (68.33%) among the three criteria, especially for long data lengths. When CCIC was applied to the MA(3) model, its average accuracy was 52.45%, still keeping higher accuracy than that of AIC and BIC for different data lengths. Hence, it was found that CCIC had high efficiency for the determination of a suitable MA(q) order.

Appendix B: Results of Monte-Carlo Experiments for ARMA Models

Using an autoregressive moving average model ARMA(p,q), the residual time series x_t can be described as:

$$x_t = u + \varphi_1 (x_{t-1} - u) + \varphi_2 (x_{t-2} - u) + \cdots + \varphi_p (x_{t-p} - u) + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{B1}$$

where u is the mean of x_t; φ_1, φ_2, …, φ_p are the autoregressive coefficients; p is the autoregressive order; θ_1, θ_2, …, θ_q are the moving average coefficients; q is the moving average order; ε_t has the same meaning as that in Equation 1.

By denoting the dependent component $\varphi_1 (x_{t-1} - u) + \varphi_2 (x_{t-2} - u) + \cdots + \varphi_p (x_{t-p} - u) - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$ as η_t, and denoting the random component $u + \varepsilon_t$ as u_t, x_t can be expressed as their sum:

$$x_t = \eta_t + u_t \tag{B2}$$

where η_t and u_t are independent, and $E(x_t) = u$, $E(u_t) = u$, and $E(\eta_t) = 0$.

Then, the same derivations as that in Equations 2–8 can also be obtained accordingly to get the following equation:

$$r^2 = 1 - \frac{{\sigma_u}^2}{{\sigma_x}^2} \tag{B3}$$

In order to establish the relationship between the correlation coefficient r and the ARMA model's parameters, the following is further derived.

Transposing Equation B1, we can get:

$$x_t - u = \varphi_1 (x_{t-1} - u) + \varphi_2 (x_{t-2} - u) + \cdots + \varphi_p (x_{t-p} - u) + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{B4}$$

Multiplying both sides of Equation B4 by $x_t - u$, and taking their expectation (E), one gets:

$$\begin{aligned} E(x_t - u)^2 = {} & \varphi_1 E\left[(x_{t-1} - u)(x_t - u)\right] + \varphi_2 E\left[(x_{t-2} - u)(x_t - u)\right] \\ & + \cdots + \varphi_p E\left[(x_{t-p} - u)(x_t - u)\right] + E\left[(x_t - u)\varepsilon_t\right] - \theta_1 E\left[(x_t - u)\varepsilon_{t-1}\right] \\ & - \theta_2 E\left[(x_t - u)\varepsilon_{t-2}\right] - \cdots - \theta_q E\left[(x_t - u)\varepsilon_{t-q}\right] \end{aligned} \tag{B5}$$

Dividing both sides of Equation B5 by σ_x^2, and considering that

$$\frac{E\left[(x_{t-i} - u)(x_t - u)\right]}{\sigma_x^2} = \rho_i,\ \rho_0 = 1 \tag{B6}$$

we can get:

$$1 - \rho_1 \varphi_1 - \rho_2 \varphi_2 - \cdots - \rho_P \varphi_P = \frac{E\left[(x_t - u)\varepsilon_t\right] - \theta_1 E\left[(x_t - u)\varepsilon_{t-1}\right] - \cdots - \theta_q E\left[(x_t - u)\varepsilon_{t-q}\right]}{\sigma_x^2} \tag{B7}$$

Multiplying both sides of Equation B7 by σ_x^2, one gets:

$$(1 - \rho_1 \varphi_1 - \rho_2 \varphi_2 - \cdots - \rho_P \varphi_P)\,\sigma_x^2 = E\left[(x_t - u)\varepsilon_t\right] - \theta_1 E\left[(x_t - u)\varepsilon_{t-1}\right] - \cdots - \theta_q E\left[(x_t - u)\varepsilon_{t-q}\right] \tag{B8}$$

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Multiplying both sides of Equation B4 by ε_t, and taking their expectation (E), one gets:

$$E\left[(x_t-u)\,\varepsilon_t\right]=\varphi_1E\left[(x_{t-1}-u)\,\varepsilon_t\right]+\varphi_2E\left[(x_{t-2}-u)\,\varepsilon_t\right]+\cdots+\varphi_pE\left[(x_{t-p}-u)\,\varepsilon_t\right]$$
$$+E\left(\varepsilon_t^2\right)-\theta_1E\left(\varepsilon_t\varepsilon_{t-1}\right)-\theta_2E\left(\varepsilon_t\varepsilon_{t-2}\right)-\cdots-\theta_qE\left(\varepsilon_t\varepsilon_{t-q}\right)=E\left(\varepsilon_t^2\right)=\sigma_\varepsilon^2 \quad \text{(B9)}$$

Multiplying both sides of Equation B4 by ε_{t-1}, and taking their expectation (E), one gets:

$$E\left[(x_t-u)\,\varepsilon_{t-1}\right]=\varphi_1E\left[(x_{t-1}-u)\,\varepsilon_{t-1}\right]+\varphi_2E\left[(x_{t-2}-u)\,\varepsilon_{t-1}\right]+\cdots+\varphi_pE\left[(x_{t-p}-u)\,\varepsilon_{t-1}\right]$$
$$+E\left(\varepsilon_t\varepsilon_{t-1}\right)-\theta_1E\left(\varepsilon_{t-1}^2\right)-\theta_2E\left(\varepsilon_{t-1}\varepsilon_{t-2}\right)$$
$$-\cdots-\theta_qE\left(\varepsilon_{t-1}\varepsilon_{t-q}\right)=-\theta_1E\left(\varepsilon_{t-1}^2\right)=-\theta_1\sigma_\varepsilon^2 \quad \text{(B10)}$$

Similarly, and so on, until multiplying both sides of Equation B4 by ε_{t-q}, and taking their expectation (E), one gets:

$$E\left[(x_t-u)\,\varepsilon_{t-q}\right]=\varphi_1E\left[(x_{t-1}-u)\,\varepsilon_{t-q}\right]+\varphi_2E\left[(x_{t-2}-u)\,\varepsilon_{t-q}\right]+\cdots+\varphi_pE\left[(x_{t-p}-u)\,\varepsilon_{t-q}\right]$$
$$+E\left(\varepsilon_t\varepsilon_{t-q}\right)-\theta_1E\left(\varepsilon_{t-1}\varepsilon_{t-q}\right)$$
$$-\theta_2E\left(\varepsilon_{t-2}\varepsilon_{t-q}\right)-\cdots-\theta_qE\left(\varepsilon_{t-q}^2\right)=-\theta_qE\left(\varepsilon_{t-q}^2\right)=-\theta_q\sigma_\varepsilon^2 \quad \text{(B11)}$$

Therefore, Equation B8 can be simplified as:

$$\left(1-\rho_1\varphi_1-\rho_2\varphi_2-\cdots-\rho_P\varphi_P\right)\sigma_x^2=\left(1+\theta_1^2+\theta_2^2+\cdots+\theta_q^2\right)\sigma_\varepsilon^2 \quad \text{(B12)}$$

Equation B12 can be re-written as:

$$\frac{\sigma_\varepsilon^2}{\sigma_x^2}=\frac{1-\rho_1\varphi_1-\rho_2\varphi_2-\cdots-\rho_P\varphi_P}{1+\theta_1^2+\theta_2^2+\cdots+\theta_q^2} \quad \text{(B13)}$$

As $\sigma_\varepsilon^2=\sigma_u{}^2$, combining Equation B3 and Equation B13, we obtain:

$$r^2=1-\frac{1-\rho_1\varphi_1-\rho_2\varphi_2-\cdots-\rho_P\varphi_P}{1+\theta_1^2+\theta_2^2+\cdots+\theta_q^2} \quad \text{(B14)}$$

When $p=0$, Equation B14 presents a MA model, and it can be written as:

$$r^2=1-\frac{1}{\left(1+\theta_1^2+\theta_2^2+\cdots+\theta_q^2\right)} \quad \text{(B15)}$$

When $q=0$, Equation B14 presents an AR model, and it can be written as:

$$r^2=\rho_1\varphi_1+\rho_2\varphi_2+\cdots+\rho_P\varphi_P \quad \text{(B16)}$$

Equation B15 and Equation B16 are the same as Equation A9 in Appendix A and Equation 14, respectively. Therefore, it is thought that the above derivations of Equation B14 are reasonable, and clearly indicate the relationship between the correlation coefficient r and the parameters for the ARMA model.

We design the Monte-Carlo experiments by taking the ARMA(1,1), ARMA(1,2), and ARMA(2,1) models as examples to compare the accuracy of CCIC with that of AIC and BIC. The relevant parameter values and the experiment steps are just the same as that for the AR and MA models.

For the ARMA(1,1) model, its expression can be written as $x_t=\varphi_1(x_{t-1}-u)-\theta_1\varepsilon_{t-1}+u_t$. Based on the Yule-Walker equations, we obtain $\rho_1=\varphi_1$ and insert into Equation B14, then r^2 can be shown as:

$$r^2=1-\frac{1-\varphi_1^2}{1+\theta_1{}^2} \quad \text{(B17)}$$

The ARMA(1,2) model is expressed as: $x_t=\varphi_1(x_{t-1}-u)-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}+u_t$. Based on the Yule-Walker equations, we obtain $\rho_1=\varphi_1$ and insert into Equation B14, then r^2 can be shown as:

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table B1
Accuracy of the Determination of ARMA Model Orders Under Different Data Lengths by AIC, BIC, and the Proposed CCIC

	ARMA(1,1)			ARMA(1,2)			ARMA(2,1)		
Data length	AIC	BIC	CCIC	AIC	BIC	CCIC	AIC	BIC	CCIC
$n = 50$	16.87%	62.33%	83.80%	11.40%	39.67%	52.47%	12.80%	44.00%	62.87%
$n = 75$	17.47%	69.47%	86.73%	14.73%	49.20%	59.27%	13.93%	53.07%	68.13%
$n = 100$	18.27%	72.53%	90.53%	16.47%	51.40%	61.60%	15.33%	57.00%	72.40%
$n = 150$	18.87%	75.13%	91.87%	16.67%	53.20%	65.20%	17.53%	59.27%	74.47%
$n = 200$	21.40%	75.73%	94.20%	18.20%	56.73%	70.27%	18.93%	60.40%	79.27%
Mean value	18.57%	71.04%	89.43%	15.49%	50.04%	61.76%	15.71%	54.75%	71.43%

$$r^2 = 1 - \frac{1 - \varphi_1^2}{1 + {\theta_1}^2 + {\theta_2}^2} \tag{B18}$$

The ARMA(2,1) model is expressed as: $x_t = \varphi_1 (x_{t-1} - u) + \varphi_2 (x_{t-2} - u) - \theta_1 \varepsilon_{t-1} + u_t$, with $\rho_1 = \frac{\varphi_1}{1-\varphi_2}$ and $\rho_2 = \frac{\varphi_1^2}{1-\varphi_2} + \varphi_2$ from the Yule-Walker equations, we can obtain:

$$r^2 = 1 - \frac{1 - \left[\frac{\varphi_1^2(1+\varphi_2)}{1-\varphi_2} + \varphi_2^2\right]}{1 + {\theta_1}^2} = \frac{{\theta_1}^2 + \frac{\varphi_1^2(1+\varphi_2)}{1-\varphi_2} + \varphi_2^2}{1 + {\theta_1}^2} \tag{B19}$$

Table B1 shows the accuracy of correctly identifying the true order of ARMA models. Similarly, it was found that CCIC had higher efficiency for the determination of suitable ARMA(p,q) model order compared to AIC and BIC, no matter which data lengths are considered.

Data Availability Statement

The data were obtained from the China Meteorological Data Sharing Service System (http://cdc.cma.gov.cn/).

Acknowledgments
The authors gratefully acknowledge the valuable comments and suggestions given by the editors and the two anonymous reviewers. This study was financially supported by the National Key Research and Development Program (No. 2019YFA0606903), the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) (No. XDA20060402), and the National Natural Science Foundation of China (No. 91547205, 41971040, 51779176).

References

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: The worldviews of AIC and BIC. *Ecology*, *95*(3), 631–636. https://doi.org/10.1890/13-1452.1

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Breakthroughs in Statistics*, 610–624. https://doi.org/10.1007/978-1-4612-0919-5_38

Asfaw, A., Simane, B., Hassen, A., & Bantider, A. (2018). Variability and time series trend analysis of rainfall and temperature in northcentral Ethiopia: A case study in Woleka sub-basin. *Weather and Climate Extremes*, *19*, 29–41. https://doi.org/10.1016/j.wace.2017.12.002

Bomblies, A., Duchemin, J. B., & Eltahir, E. A. B. (2008). Hydrology of malaria: Model development and application to a Sahelian village. *Water Resources Research*, *44*(12), W12445. https://doi.org/10.1029/2008wr006917

Castillo, A., Castelli, F., & Entekhabi, D. (2015). An entropy-based measure of hydrologic complexity and its applications. *Water Resources Research*, *51*(7), 5145–5160. https://doi.org/10.1002/2014wr016035

Debele, S. E., Bogdanowicz, E., & Strupczewski, W. G. (2017). The impact of seasonal flood peak dependence on annual maxima design quantiles. *Hydrological Sciences Journal*, *62*(10), 1603–1617. https://doi.org/10.1080/02626667.2017.1328558

Fathian, F., Dehghan, Z., Bazrkar, M. H., & Eslamian, S. (2016). Trends in hydrological and climatic variables affected by four variations of the Mann-Kendall approach in Urmia Lake basin, Iran. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, *61*(5), 892–904. https://doi.org/10.1080/02626667.2014.932911

Fishler, E., Grosmann, M., & Messer, H. (2002). Detection of signals by information theoretic criteria: General asymptotic performance analysis. *IEEE Transactions on Signal Processing*, *50*, 1027–1036. https://doi.org/10.1109/78.995060

Fort, A., Manfredi, C., & Rocchi, S. (1995). Adaptive SVD-based AR model order determination for time-frequency analysis of Doppler ultrasound signals. *Ultrasound in Medicine and Biology*, *21*(6), 793–805. https://doi.org/10.1016/0301-5629(95)00021-i

Franke, J. (1985). A Levinson Durbin recursion for autoregressive-moving average processes. *Biometrika*, *72*(3), 573–581. https://doi.org/10.1093/biomet/72.3.573

Gaonkar, B., Beckett, J., Attiah, M., Ahn, C., Edwards, M., & Wilson, B. (2021). Eigenrank by committee: Von-Neumann entropy based data subset selection and failure prediction for deep learning based medical image segmentation. *Medical Image Analysis*, *67*. https://doi.org/10.1016/j.media.2020.101834

Han, X., Hsieh, C.-S., & Lee, L.-f. (2017). Estimation and model selection of higher-order spatial autoregressive model: An efficient Bayesian approach. *Regional Science and Urban Economics*, *63*, 97–120. https://doi.org/10.1016/j.regsciurbeco.2016.12.003

Herrick, J. (2014). Signal to noise. In *The Antioch review*, (pp. 705). Antioch Review, Incorporated, Yellow Springs.

Hodgkins, G. A., Dudley, R. W., Nielsen, M. G., Renard, B., & Qi, S. L. (2017). Groundwater-level trends in the U.S. glacial aquifer system, 1964–2013. *Journal of Hydrology*, *553*, 289–303. https://doi.org/10.1016/j.jhydrol.2017.07.055

Hossain, M. B., Moon, J., & Chon, K. (2020). *Estimation of ARMA model order via Artificial Neural Network for modeling Physiological systems* (Vol. 8). IEEE Access. https://doi.org/10.1109/access.2020.3029756

Huang, J. T., Yong, W. A., & Hong, L. (2016). Generalization of the Kullback-Leibler divergence in the Tsallis statistics. *Journal of Mathematical Analysis and Applications*, *436*(1), 501–512. https://doi.org/10.1016/j.jmaa.2015.12.008

Iliopoulou, T., Papalexiou, S. M., Markonis, Y., & Koutsoyiannis, D. (2018). Revisiting long-range dependence in annual precipitation. *Journal of Hydrology*, *556*, 891–900. https://doi.org/10.1016/j.jhydrol.2016.04.015

Jiang, C., Xiong, L. H., Xu, C. Y., & Guo, S. L. (2015). Bivariate frequency analysis of nonstationary low-flow series based on the time-varying copula. *Hydrological Processes*, *29*(6), 1521–1534. https://doi.org/10.1002/hyp.10288

Jiang, L., Li, N. N., & Zhao, X. (2017). Scaling behaviors of precipitation over China. *Theoretical and Applied Climatology*, *128*(1–2), 63–70. https://doi.org/10.1007/s00704-015-1689-2

Jirak, M. (2012). Simultaneous confidence bands for Yule-Walker estimators and order selection. *Annals of Statistics*, *40*(1), 494–528. https://doi.org/10.1214/11-aos963

Khan, D. M., Yahya, N., Kamel, N., & Ieee (2021). *Optimum order selection criterion for autoregressive models of Bandlimited EEG signals*, paper presented at *IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES) - Leading Modern Healthcare Technology Enhancing Wellness*, Electr Network, Mar 01-03. https://doi.org/10.1109/iecbes48179.2021.9398836

Khorshidi, S., Karimi, M., & Nematollahi, A. R. (2011). New autoregressive (AR) order selection criteria based on the prediction error estimation. *Signal Processing*, *91*(10), 2359–2370. https://doi.org/10.1016/j.sigpro.2011.04.021

Koirala, S. R., Gentry, R. W., Mulholland, P. J., Perfect, E., Schwartz, J. S., & Sayler, G. S. (2011). Persistence of hydrologic variables and reactive stream solute concentrations in an east Tennessee watershed. *Journal of Hydrology*, *401*(3–4), 221–230. https://doi.org/10.1016/j.jhydrol.2011.02.022

Kong, X. M., Huang, G. H., Fan, Y. R., & Li, Y. P. (2015). Maximum entropy-Gumbel-Hougaard copula method for simulation of monthly streamflow in Xiangxi river, China. *Stochastic Environmental Research and Risk Assessment*, *29*(3), 833–846. https://doi.org/10.1007/s00477-014-0978-0

Koutsoyiannis, D. (2014). Entropy: From thermodynamics to hydrology. *Entropy*, *16*(3), 1287–1314. https://doi.org/10.3390/e16031287

Li, D., & Jayaweera, S. K. (2017). Uncertainty modeling and price-based demand response scheme design in smart grid. *IEEE Systems Journal*, *11*(3), 1743–1754. https://doi.org/10.1109/jsyst.2014.2369451

Lin, L. C., Huang, P. H., & Weng, L. J. (2017). Selecting path models in SEM: A comparison of model selection criteria. *Structural Equation Modeling-a Multidisciplinary Journal*, *24*(6), 855–869. https://doi.org/10.1080/10705511.2017.1363652

Liu, Y., Huang, Z. P., Su, S. J., & Liu, C. W. (2012). AR model whitening and signal detection based on GLD algorithm in the non-Gaussian reverberation. *Applied Acoustics*, *73*(10), 1045–1051. https://doi.org/10.1016/j.apacoust.2012.05.002

Machiwal, D., & Jha, M. K. (2012). Current status of time series analysis in hydrological Sciences, In *Hydrologic time series analysis: Theory and Practice*, (pp. 96–136). Springer Netherlands. https://doi.org/10.1007/978-94-007-1861-6_6

Maesono, Y. (2005). Asymptotic representation of ratio statistics and their mean squared errors. *Journal of the Japan Statistical Society*, *35*. https://doi.org/10.14490/jjss.35.73

Markonis, Y., & Koutsoyiannis, D. (2016). Scale-dependence of persistence in precipitation records. *Nature Climate Change*, *6*(4), 399–401. https://doi.org/10.1038/nclimate2894

Markonis, Y., Moustakis, Y., Nasika, C., Sychova, P., Dimitriadis, P., Hanel, M., et al. (2018). Global estimation of long-term persistence in annual river runoff. *Advances in Water Resources*, *113*, 1–12. https://doi.org/10.1016/j.advwatres.2018.01.003

Militino, A. F., Moradi, M., & Ugarte, M. D. (2020). On the performances of trend and change-point detection methods for remote sensing data. *Remote Sensing*, *12*(6). https://doi.org/10.3390/rs12061008

Moon, J., Hossain, M. B., & Chon, K. H. (2021). *AR and ARMA model order selection for time-series modeling with ImageNet classification* (Vol. 183). Signal Processing. https://doi.org/10.1016/j.sigpro.2021.108026

Moravej, M., & Khalili, K. (2015). Hydrological time series analysis and modelling using statistical tests and linear time series models (case study: West Azerbaijan province of Iran). *International Journal of Hydrology Science and Technology*, *5*, 349–371. https://doi.org/10.1504/ijhst.2015.072638

Nagakura, D. (2009). Testing for coefficient stability of AR(1) model when the null is an integrated or a stationary process. *Journal of Statistical Planning and Inference*, *139*(8), 2731–2745. https://doi.org/10.1016/j.jspi.2008.12.009

Paschalis, A., Molnar, P., & Burlando, P. (2012). Temporal dependence structure in weights in a multiplicative cascade model for precipitation. *Water Resources Research*, *48*(1), W01501. https://doi.org/10.1029/2011wr010679

Peng, X. Y., Wang, M., & Liu, C. D. (2009). *AR model parameters*, *41*, 12–16. Adaptive estimation method of.

Rajsekhar, D., Singh, V. P., & Mishra, A. K. (2015). Multivariate drought index: An information theory based approach for integrated drought assessment. *Journal of Hydrology*, *526*, 164–182. https://doi.org/10.1016/j.jhydrol.2014.11.031

Ryberg, K. R., Hodgkins, G. A., & Dudley, R. W. (2020). Change points in annual peak streamflows: Method comparisons and historical change points in the United States. *Journal of Hydrology*, *583*. https://doi.org/10.1016/j.jhydrol.2019.124307

Sagarika, S., Kalra, A., & Ahmad, S. (2014). Evaluating the effect of persistence on long-term trends and analyzing step changes in streamflows of the continental United States. *Journal of Hydrology*, *517*, 36–53. https://doi.org/10.1016/j.jhydrol.2014.05.002

Sang, Y. F., Singh, V. P., Hu, Z., Xie, P., & Li, X. (2018). Entropy-aided evaluation of meteorological droughts over China. *Journal of Geophysical Research-Atmospheres*, *123*(2), 740–749. https://doi.org/10.1002/2017jd026956

Sang, Y. F., Singh, V. P., Wen, J., & Liu, C. M. (2015). Gradation of complexity and predictability of hydrological processes. *Journal of Geophysical Research – Atmospheres*, *120*(11), 5334–5343. https://doi.org/10.1002/2014jd022844

Sang, Y. F., Sivakumar, B., & Zhu, Y. (2021). Uniform discrete wavelet spectrum for detection of hydrologic variability at multiple timescales. *Journal of Hydro-Environment Research*, *35*, 31–37. https://doi.org/10.1016/j.jher.2021.01.005

Sang, Y. F., Sun, F., Singh, V. P., Xie, P., & Sun, J. (2018). A discrete wavelet spectrum approach to identifying non-monotonic trend pattern of hydroclimate data. *Hydrology and Earth System Sciences*, *22*, 757–766. https://doi.org/10.5194/hess-22-757-2018

Serinaldi, F., & Kilsby, C. G. (2016). The importance of prewhitening in change point analysis under persistence. *Stochastic Environmental Research and Risk Assessment*, *30*(2), 763–777. https://doi.org/10.1007/s00477-015-1041-5

Singh, V. P. (2013). *Environment theory and its application in Enviromental and water engineering*. Wiley-Blackwell.

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

10.1029/2021WR031606

Slawski, M., & Hein, M. (2015). Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random fields. *Linear Algebra and its Applications*, *473*, 145–179. https://doi.org/10.1016/j.laa.2014.04.020

Tan, X. Z., Gan, T. Y., & Shao, D. G. (2017). Effects of persistence and large-scale climate anomalies on trends and change points in extreme precipitation of Canada. *Journal of Hydrology*, *550*, 453–465. https://doi.org/10.1016/j.jhydrol.2017.05.028

Wang, Q. S., Li, H. P., Lin, J. Y., & Zhang, C. X. (2018). Future trend forecast by empirical wavelet transform and autoregressive moving average. *Sensors*, *18*(8). https://doi.org/10.3390/s18082621

Wilson, E. M. (2014). Engineering hydrology. In H. Herrmann & H. Bucksch (Eds.), *Dictionary Geotechnical Engineering/Wörterbuch GeoTechnik: English-German/Englisch-Deutsch*, (pp. 472).

Wu, Z. Y., Xie, P., Sang, Y. F., Chen, J., Ke, W., Zhao, J. Y., & Zhao, Y. X. (2019). Moving correlation coefficient-based method for jump points detection in hydroclimate time series. *Stochastic Environmental Research and Risk Assessment*, *33*(10), 1751–1764. https://doi.org/10.1007/s00477-019-01727-6

Xie, P., Gu, H. T., Sang, Y. F., Wu, Z. Y., & Singh, V. P. (2019). Comparison of different methods for detecting change points in hydroclimatic time series. *Journal of Hydrology*, *577*, 123973. https://doi.org/10.1016/j.jhydrol.2019.123973

Xie, P., Wu, L. Q., Sang, Y. F., Chan, F. K. S., Chen, J., Wu, Z. Y., & Li, Y. Q. (2021). Correlation-aided method for identification and gradation of periodicities in hydrologic time series. *Geoscience Letters*, *8*(1). https://doi.org/10.1186/s40562-021-00183-x

Xie, P., Wu, Z. Y., Sang, Y. F., Gu, H. T., Zhao, Y. X., & Singh, V. P. (2018). Evaluation of the significance of abrupt changes in precipitation and runoff process in China. *Journal of Hydrology*, *560*, 451–460. https://doi.org/10.1016/j.jhydrol.2018.02.036

Xie, P., Zhao, Y. X., Sang, Y. F., Gu, H. T., Wu, Z. Y., & Singh, V. P. (2018). Gradation of the significance level of trends in precipitation over China. *Hydrology Research*, *49*(6), 1890–1901. https://doi.org/10.2166/nh.2018.187

Xie, Y. Y., Huang, Q., Chang, J. X., Liu, S. Y., & Wang, Y. M. (2016). Period analysis of hydrologic series through moving-window correlation analysis method. *Journal of Hydrology*, *538*, 278–292. https://doi.org/10.1016/j.jhydrol.2016.04.024

Yang, Y. Q., Xu, P., Yang, G. T., Chen, L., & Li, J. B. (2021). BIC-based data-driven rail track deterioration adaptive piecewise modeling framework. *Frontiers in Materials*, *8*. https://doi.org/10.3389/fmats.2021.620484

Ye, X. C., Xu, C. Y., Li, X. H., & Zhang, Q. (2018). Investigation of the complexity of streamflow fluctuations in a large heterogeneous lake catchment in China. *Theoretical and Applied Climatology*, *132*(3–4), 751–762. https://doi.org/10.1007/s00704-017-2126-5

Yue, S., & Pilon, P. (2003). Interaction between deterministic trend and autoregressive process. *Water Resources Research*, *39*(4), 1077. https://doi.org/10.1029/2001wr001210

Yue, S., Pilon, P., & Cavadias, G. (2002). Power of the Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of Hydrology*, *259*(1–4), 254–271. https://doi.org/10.1016/s0022-1694(01)00594-7

Zar, J. H. (1972). Significance testing of the Spearman rank correlation coefficient. *Journal of the American Statistical Association*, *67*(339), 578–580. https://doi.org/10.1080/01621459.1972.10481251

Zhang, P. (2007). On the convergence rate of model selection criteria. *Communications in Statistics - Theory and Methods*, *22*(10), 2765–2775.

Zhang, Q., Singh, V. P., Li, K., & Li, J. (2014). Trend, periodicity and abrupt change in streamflow of the East River, the Pearl River basin. *Hydrological Processes*, *28*(2), 305–314. https://doi.org/10.1002/hyp.9576

Zhao, J., Xie, P., Zhang, M., Sang, Y.-F., Chen, J., & Wu, Z. (2018). Nonstationary statistical approach for designing LNWLs in inland waterways: A case study in the downstream of the Lancang river. *Stochastic Environmental Research and Risk Assessment*, *32*(11), 3273–3286. https://doi.org/10.1007/s00477-018-1606-1

19447973, 2022, 7, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2021WR031606 by South University Of Science, Wiley Online Library on [07/12/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License