Understanding the impacts induced by cut-off thresholds and likelihood measures on confidence interval when applying GLUE approach

Zhenya Li\textsuperscript{a}, Tao Yang\textsuperscript{a*}, Na Zhang\textsuperscript{e}, Yandong Zhang\textsuperscript{d}, Jiahu Wang\textsuperscript{a}, Chong-Yu Xu\textsuperscript{b}, Pengfei Shi\textsuperscript{a}, Youwei Qin\textsuperscript{a}

\textsuperscript{a} State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Center for Global Change and Water Cycle, Hohai University, Nanjing 210098, China;

\textsuperscript{b} Department of Geosciences, University of Oslo, P.O. Box 1047 Blindern, 0316 Oslo, Norway;

\textsuperscript{c} CSIRO Digital Productivity Flagship, Leeuwin Centre, 65 Brockway Road Floreat, WA 6014, Australia;

\textsuperscript{d} China Water Northeastern Investigation, Design & Research Co., Ltd., Changchun130021, Jilin, China;

\textsuperscript{e} Bureau of Hydrology Changjiang Water Resources Commission, 93 Dama Road, Wuhan 210000, China;

Declarations
Data Availability
None.

Animal Research (Ethics)
Not applicable

Consent to Participate (Ethics)
Not applicable

Consent to Publish (Ethics)
Not applicable

*Corresponding author:
Dr. Tao Yang, Professor
State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering
Center for Global Change and Water Cycle
Hohai University, Nanjing 210098
The People’s Republic of China
E-mail: tao.yang@hhu.edu.cn
Author contribution
All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Zhenya Li, Tao Yang, Chong-Yu Xu, Pengfei Shi and Youwei Qin. Methodology was proposed by Zhenya Li, Na Zhang, Yandong Zhang and Jiahu Wang. The first draft of the manuscript was written by Zhenya Li and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Conflicts of interest/Competing interests
None.

Code availability
Relevant software applications and custom codes that support the findings of this work are available from the corresponding author upon reasonable request.

Funding
The work was jointly supported by a grant from the Fundamental Research Funds for the Central Universities (B200202134), China Postdoctoral Science Foundation (2020M671322), grants from the National Natural Science Foundation of China (51809072, 51879068), grants from the National Key R&D Program of China (2018YFC150806, 2018YFC0407900, 2018YFC1508602, 2019YFC1510700).

Abstract: Generalized likelihood uncertainty estimate (GLUE) approach are heavily affected by the choices of cut-off thresholds and likelihood measures. This work attempts to study the potential mechanisms behind the impacts induced by cut-off thresholds and likelihood measures on confidence interval obtained by GLUE. A theoretical analysis on typical likelihood measures reveals that the error model of likelihood measure has essential impacts on the sampling processes of GLUE. Likelihood measures based on a same error model are mathematically transferrable, leading to an identical population of acceptable parameter sets. A case study is conducted by applying GLUE to uncertainty analysis on daily flows simulated by HBV model for the source region of the Yellow River basin. Seven interval indicators are adopted to describe the geometric features of confidence intervals, which are integrated into a comprehensive score for an overall assessment by multiple
attribute decision making (MADM) framework. Results indicate that 1) With an increase of cut-off threshold, confidence interval widens in low-level flow sections, moves upward in recession phases of medium-level flow sections whereas narrows in high-level flow sections. Trade-off mechanism amongst widening, moving and narrowing trends is a potential reason behind the variations of interval indicators with cut-off threshold. 2) Much higher similarities in confidence intervals can be detected for likelihood measures based on a same error model than those based on different error models; 3) increasing cut-off threshold highlights the impacts induced by the error models of likelihood measures, whereas weakens the impacts induced by the formulas of likelihood measures.

**Keywords:** Generalized likelihood uncertainty estimate (GLUE) approach; confidence interval; likelihood measures; cut-off threshold; error model
1. Introduction

Hydrological simulation is challenged by the uncertainty of model structure, model parameter and input datasets (Beven and Binley 1992). Such uncertainty can be evaluated by a variety of approaches such as the stochastic response surface method (Cryer and Applequist 2003), Taylor expansion-based methods (Mo et al. 2017) and Monte Carlo (MC) based methods (Sahoo 2018). Of all these methods, Zheng and Keller (2007) deemed MC-based methods to be the most appropriate. Therein, generalized likelihood uncertainty estimate (GLUE) approach is one of the most extensively-used MC-based methods (e.g. He et al. 2009; Mirzaei et al. 2015; Sun et al. 2016). Apart from simplicity, GLUE exhibits the following advantages: 1) it accounts for all uncertainty sources explicitly or implicitly; 2) it is logically simple with no strict error assumptions (Houska et al. 2014); and 3) it is not sensitive to the discontinuous parameter surface of complex hydrological models characterized by tens of parameters (Zheng and Keller 2007).

As based on the concept of equifinality, GLUE accepts a series of comparatively reliable parameter sets rather than a single optimal parameter set (Beven 2006). At a time point, a hydrological model can generate a series of outputs by inputting accepted parameter sets, which can be used to obtain a posterior distribution by Bayesian theory. The p% (e.g. 90%, 95%) confidence interval of posterior distribution is often employed for a quantitative analysis on the uncertainty of model outputs (Christensen 2004). However, the confidence interval obtained by GLUE (shorten as GLUE interval in the following sections) was sometimes reported to be too wide (Lamb et al. 1998), too narrow (Jensen 2003) or even offset against observed flows (Freer et al. 1996). The instability is caused mainly by the settings of subjective conditions in GLUE, such as the number of samples (Viola et al. 2009), the ranges of model parameters (Freni et al. 1996), cut-off thresholds (Freni et al. 2008) and likelihood measures (Romanowicz and Beven 2006).

Previous works have taken some strategies of relieving the impacts induced by the number of samples and the ranges of model parameters. For example, Viola et al. (2009) stated that re-increasing the number of samples has few impacts on 90% GLUE intervals when the number of samples has been very large (e.g. > 1,000,000). Blasone et al. (2008) recommended the ranges slightly fluctuating around optimal parameter set to weaken the impacts induced by the ranges of model parameters. Thus, hydrologists paid attentions mainly to the impacts of the rest two subjective conditions, namely cut-off threshold and likelihood measure (e.g. Montanari 2005; Jung and Merwade 2011; Li et al. 2011; Houska et al. 2014). Likelihood measure is used to describe the closeness between model outputs and observed flows (Singh et al. 2010). A candidate parameter set is considered acceptable when it obtains a likelihood value of larger than a pre-defined cut-off threshold.
Freni et al. (2008, 2009) studied the impacts induced by cut-off threshold and likelihood measure by applying GLUE to an integrated urban-drainage model. Alazzy et al. (2015) compared a range of widely-used likelihood measures in the uncertainty analysis of XAJ rainfall-runoff model. He et al. (2010) urged that an arbitrary choice of likelihood measure or cut-off threshold might result in unreasonable model outputs. It can, however, be found that previous works focused mainly on the variations in the characteristics of GLUE intervals caused by different settings of subjective conditions, yet rarely investigating the potential mechanism behind.

The features of GLUE intervals are seriously affected by the settings of subjective conditions. However, a GLUE interval can hardly be assessed quantitatively because it is a combination of confidence intervals at all time steps rather than a single value. To solve this problem, scholars have proposed a number of interval indicators to characterize the various geometric features of GLUE intervals (Blasone et al. 2008; Olsson and Lindstrom 2008; Jin et al. 2010). Zhang et al. (2009) denoted the ratio of observed flows covered by GLUE interval as CR. Kasiviswanatan and Sudheer (2016) said that a CR close to a pre-defined confidence level is desirable, which is however hard to be achieved in most practical cases (Beven and Freer 2001). Zhang et al. (2015) believed that an excessively wide GLUE interval conveys little beneficial information. He introduced band-width (B) and relative band-width (RB) to measure the sharpness degree of a GLUE interval. Xiong et al. (2009) created four interval indicators related to the symmetry degree and deviation amplitude of GLUE interval. Above indicators evaluate the quality of a GLUE interval from the perspective of various geometric features, but are often found to be conflicted with each other. Lots of works have documented that a large coverage ratio, narrow band-width and high symmetry degree cannot be achieved simultaneously (Chen et al. 2013; Xiong et al. 2009). Therefore, there is an urgent necessity to carry out a multiple-indicator assessment on the overall quality of GLUE intervals.

Based on what’s stated above, this work attempts to study the potential mechanisms behind the impacts induced by cut-off threshold and likelihood measure on GLUE intervals. Several typical likelihood measures are employed for comparisons and classified into four groups based on error model. A theoretical analysis is conducted to discuss how the error model of likelihood measure affects the mathematic relationships among likelihood measures and the sampling processes of GLUE. A number of interval indicators are adopted to characterize the various geometric features of GLUE intervals. Then, these indicators are integrated into a comprehensive score by a multiple attribute decision making (MADM) framework for an assessment on the overall quality of GLUE intervals. Section 2 gives a brief description of GLUE, cut-off threshold, likelihood
measure, interval indicator and MADM framework. Particularly, the impacts induced by the error model of likelihood measures on mathematic relationships among likelihood measures and the sampling processes of GLUE are discussed in detail. Section 3 describes the study area (i.e. the source region of the Yellow River basin) of our interest and the HBV hydrological model used for daily flow simulation. Section 4 adopts three numerical experiments to analyze the potential mechanism behind the impacts induced by cut-off thresholds and likelihood measures on GLUE intervals. Finally, section 5 summarizes the conclusions of this work.

2. Methodology

2.1 Generalized likelihood uncertainty estimate (GLUE) approach

2.1.1 General processes of GLUE

Given a hydrological model and input datasets, the processes of generating 90% confidence interval by GLUE are presented as follows:

Step 1: Decide the likelihood measure, cut-off threshold, the number of samples \(N_{\text{Sampling}}\), and the prior distributions of model parameters;

Step 2: Repeat (1) to (3) for \(N_{\text{Sampling}}\) times;

(1) Sample a candidate parameter set from the prior distributions of model parameters by Monte Carlo sampling method;

(2) Run the hydrological model with sampled candidate parameter set;

(3) Calculate the likelihood value between observed flows and model outputs based on the formula of likelihood measure;

Step 3: Accept a candidate parameter set when likelihood value exceeds cut-off threshold;

Step 4: Run hydrological model with all accepted parameters sets and estimate the posterior probability density (PPD) distribution of model outputs at each time point by Bayesian theory;

Step 5: Obtain 90% GLUE interval according to the 5th and 95th percentiles of PPD distributions at all time points.

Refer to Blasone et al. (2008) for details of estimating PPD distributions by Bayesian theory in step 4.

2.1.2 Cut-off threshold

Cut-off threshold \(Tr\) determines how large likelihood value is acceptable in the step 4 of GLUE. \(Tr\) is generally defined as a value (e.g. 0.8, 0.9) (Sun et al. 2016) or a percentage (e.g. 80%, 90%) (Li et al. 2010). A candidate parameter set is accepted only when its likelihood value is greater than \(Tr\). The percentage form
of $p\%$ is adopted in this work because the likelihood values calculated by different likelihood measures fall into different ranges and are thus incomparable (Please see Table 1). $Tr=p\%$ means that candidate parameter sets with a likelihood value of the top $p\%$ are accepted while those with a likelihood value of the rest $(1-p)\%$ are rejected.

### 2.1.3 Theoretical analysis on the formula and error model of likelihood measure

Table 1 presents the details of nine typical likelihood measures, including the Nash-Sutcliffe coefficient of efficiency (CE), inverse of residual variance (IRV), exponential likelihood (EL), normalized absolute error (NAE), inverse mean absolute error (IMAE), Chiew and McMahon (CM), logarithmic EL (LEL), mean absolute relative error (MRE) and exponential mean absolute relative error (EMRE).

These likelihood measures can be categorized into five groups according to error models. For example, the formulas of CE, IRV and EL (i.e. L1, L2 and L3 in Table 1) have a same component of $\sum_{i=1}^{n}(Eq_i - Q_i)^2$, which is a typical error model called square error (SE). It can be inferred that the three likelihood measures can be transformed mathematically into each other by the following formulas:

\[ CE = 1 - \frac{1}{\text{Var}(Q)} IRV(N_1)^{-1/N_1} \]  
\[ CE = 1 + \frac{1}{N_2} \ln[EL(N_2)] \]  
\[ EL(N_2) = \exp\left(-\frac{N_2}{\text{Var}(Q)} IRV(N_1)^{-1/N_1}\right) \]

where:

\[ \text{Var}(Q) = \sum_{i=1}^{n}(Q_i - \overline{Q})^2 \]

where $Q_i$ is the observed flow at the $i^{th}$ time point, $\overline{Q}$ is the mean of observed flows while $N_1$ and $N_2$ are the factor exponents of IRV and EL, respectively. Due to above reasons, one can classify CE, IRV and EL into a group called square error (SE) group.

Likewise, NAE and IMAE are categorized into absolute error (AE) group, CM into square root error (SRE) group, LEL into square logarithmic error (SLE) group, and MRE and EMRE into absolute relative error (ARE) group. The likelihood measures within a group (apart from SE group) are also mathematically transformable. As to NAE and IMAE in AE group, there is:
Table 1 Nine likelihood measures as well as their formulas, ranges and error models

| Name                          | Formula                                                                 | Range          | Error model          |
|-------------------------------|-------------------------------------------------------------------------|----------------|----------------------|
| L1 Nash-Sutcliffe Coefficient of Efficiency (CE) | $CE = 1 - \frac{\sum_{i=1}^{n}(E_{Qi} - Q_{i})^2}{\sum_{i=1}^{n}(Q_{i} - \bar{Q})^2}$ | $(-\infty,1]$ |                      |
| L2 Inverse of residual variance (IRV) | $IRV(N) = (\frac{\sum_{i=1}^{N}(E_{Qi} - Q_{i})^2}{\sum_{i=1}^{N}(Q_{i} - \bar{Q})^2})^{-N}$ | $(0, +\infty)$ | Square error (SE)    |
| L3 Exponential likelihood (EL) | $EL(N) = \exp(-N\sum_{i=1}^{n}(E_{Qi} - Q_{i})^2/\sum_{i=1}^{n}(Q_{i} - \bar{Q})^2)$ | $(0, 1]$       |                      |
| L4 Normalized absolute error (NAE) | $NAE = 1 - \frac{\sum_{i=1}^{n}|E_{Qi} - Q_{i}|/\sum_{i=1}^{n}|Q_{i} - \bar{Q}|}{0}$ | $(-\infty, 1]$ | Absolute error (AE)  |
| L5 Inverse mean absolute error (IMAE) | $IMAE = (\frac{1}{n}\sum_{i=1}^{n}|E_{Qi} - Q_{i}|)^{-1}$ | $(0, +\infty)$ |                      |
| L6 Chiew and McMahon (CM) | $MC = 1 - \frac{\sum_{i=1}^{n}(\sqrt{E_{Qi}} - \sqrt{Q_{i}})^2}{\sum_{i=1}^{n}(\sqrt{Q_{i}} - \bar{Q})^2}$ | $(-\infty, 1]$ | Square root error (SRE) |
| L7 Logarithmic EL (LEL) | $LEL(N) = \exp(-N\sum_{i=1}^{n}(\ln(E_{Qi}) - \ln(Q_{i}))^2/\sum_{i=1}^{n}(\ln(Q_{i}) - \ln(\bar{Q}))^2)$ | $(0, 1]$       | Square logarithmic error (SLE) |
| L8 Mean absolute relative error (MRE) | $MRE = 1 - \frac{\sum_{i=1}^{n}|E_{Qi} - Q_{i}|/Q_{i}}{n}$ | $(-\infty, 1]$ | Absolute relative error (ARE) |
| L9 Exponential mean absolute relative error (EMRE) | $EMRE = \exp(-\frac{\sum_{i=1}^{n}|E_{Qi} - Q_{i}|/Q_{i}}{n})$ | $(0, 1]$       |                      |

Note 1: $E_{Qi}$ and $Q_{i}$ is the simulated and observed flows at the $i^{th}$ time point, respectively; $\bar{Q}$ is the mean of observed flows and $n$ is the number of time points.

Note 2: Exponent $N$ in some likelihood measures is a factor exponent.
\[
IMAЕ = \left( \frac{1}{n} \sum_{i=1}^{n} |Q_i - \overline{Q}| \right) (1 - NAE)
\]  

and, as to MRE and EMRE in ARE group, there is:

\[
MRE = 1 + \ln(EMRE)
\]

The grouping of likelihood measures and the formulas of error models are also presented in Table 1.

To explore the impacts induced by the error models of likelihood measures on the processes of GLUE, the likelihood measures of CE, IRV and EL in SE group are used as examples. By recalling equations (1) to (3), it can be found that CE, IRV and EL exhibit strict positive relationships with each other when \(N_1\) and \(N_2\) are fixed. Denote the likelihood value calculated by a likelihood measure \(A\) and a parameter set \(\theta\) as \(V(A, \theta)\). If \(V(CE, \theta_1) > V(CE, \theta_2)\), there must be \(V(IRV, \theta_1) > V(IRV, \theta_2)\) and \(V(EL, \theta_1) > V(EL, \theta_2)\). This also indicates that the rankings of candidate parameter sets in the order of likelihood values are identical for CE, IRV and EL. If \(Tr\) is a fixed percentage of \(p\%\), the populations of accepted parameter sets will be identical in the step 3 of GLUE. On the other hand, the differences in the formulas of CE, IRV and EL result in different likelihood values for a same parameter set, therefore yielding different PPD distributions of model outputs in step 4 and different 90% GLUE intervals in step 5. It is worthy of noting that above analysis applies only to likelihood measures based on a same error model but not to those based on different error models. The essential reason is that no strict mathematical relationships as equations (1) to (7) can be derived among likelihood measures based on different error models. Given a fixed \(p\%\), the populations of accepted parameter sets are different in step 3, leading to different PPD distributions of model outputs in step 4 and different 90% GLUE intervals in step 5.

As a whole, both the error models and formulas of likelihood measures exhibit essential impacts on the processes of GLUE and associated GLUE intervals. The error models of likelihood measures determine the population of accepted parameter sets in the step 3 of GLUE, whereas the formulas of likelihood measures determine the characteristics of PPD distributions in the step 4 of GLUE. Detailed impacts of the two factors on GLUE intervals are discussed by three cases in section 4.

### 2.2 Quantification on the geometric features of GLUE intervals

#### 2.2.1 Typical interval indicators
## Table 2 Typical interval indicators used for characterizing the various geometric characteristics of GLUE intervals

| Property          | Name                                      | Formula                                                      | References                      | Description of variables                                                                 |
|-------------------|-------------------------------------------|--------------------------------------------------------------|---------------------------------|--------------------------------------------------------------------------------------------|
| **Coverage**      | Coverage ratio (CR)                       | \[ CR = \frac{n_c}{n} \]                                    | Zhang et al. (2009); Kasiviswanatan and Sudheer (2016) | \( n \) is the number of observed flows; \( n_c \) is the number of observed flows covered by GLUE interval; |
|                   | Average band-width (B)                    | \[ B = \frac{1}{n} \sum_{i=1}^{n} (q_i^n - q_i^l) \]        | Zhang et al. (2015)             |                                                                                             |
|                   | Average relative band-width (RB)          | \[ RB = \frac{1}{n} \sum_{i=1}^{n} \frac{(q_i^n - q_i^l)}{Q_i} \] |                                  |                                                                                             |
| **Symmetry**      | Average asymmetry degree 1 (S)            | \[ S = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{q_i^n - Q_i}{q_i^l - q_i^u} - 0.5 \right| \] | Xiong et al. (2009); Chen et al. (2013) | \( q_i^n \) and \( q_i^l \) is the upper and lower bounds of intervals at the \( i^{th} \) time point; |
|                   | Average asymmetry degree 2 (Ts)           | \[ Ts = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{q_i^n - Q_i}{q_i^l - q_i^u} \right) \left| \frac{(q_i^n - q_i^l)}{(q_i^n - q_i^u)} \right|^{1/3} \] |                                  |                                                                                             |
| **Deviation**     | Average deviation amplitude (D)           | \[ D = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{q_i^n - q_i^l}{2Q_i} - 1 \right| \] | Xiong et al. (2009); Chen et al. (2013) |                                                                                             |
| amplitude         | Average relative deviation amplitude (RD) | \[ RD = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{q_i^n - q_i^l}{2Q_i} - 1 \right| \] |                                  |                                                                                             |
Table 2 lists 7 interval indicators typically used for characterizing different geometric features of GLUE intervals. The indicators are further divided into four types related to coverage ratio, band-width, symmetry and deviation amplitude. Some indicators are divided into the same type, which focus on different aspects of a certain interval feature. For example, both B and RB characterize the sharpness degree of a GLUE interval, the former of which considers flow magnitude and the latter removes the impacts of flow magnitude. Please refer to the cited references listed in Table 2 for the details of above indicators. Theoretically, an ideal GLUE interval shall have a large coverage ratio of observed flows (i.e. large CR), narrow band-width (i.e. small B and RB), high symmetry degree (i.e. low S and Ts) and short distance from interval centers to observed flows (i.e. small D and RD) (Xiong et al. 2009). It is, however, reported in many works that these qualities cannot be achieved simultaneously (e.g. Chen et al. 2013; Xiong et al. 2009). The overall assessment of all interval indicators is a typical multiple attribute decision making (MADM) problem, which has rarely been studied before.

2.2.2 Multiple attribute decision making (MADM) framework

Given a Tr and a likelihood measure, 90% GLUE interval can be generated by the steps presented in section 2.1.1. Then, all interval indicators listed in Table 2 can be calculated and stored into a vector of \( S = (a_1, a_2, \ldots, a_i, \ldots, a_7) \) where \( a_i \) is the value of the \( j \)th interval indicator \( (j=1, 2, \ldots, 7) \). Vector \( S \) is called an “event”. Different combinations of Tr s and likelihood measures yield different 90% GLUE intervals and associated event vectors. If there are \( n \) different events, an event matrix of \( A = [a_{ij}]_{n \times m} \) can be constructed with \( a_{ij} \) being the value of the \( j \)th indicator in the \( i \)th event \( (i=1, 2, \ldots, n; j=1, 2, \ldots, 7) \).

MADM framework has been widely adopted to identify the best event from multiple ones (Hwang and Yoon 1981; Kelemenis and Askounis 2010). This work employs the MADM framework proposed by Li et al. (2018) to integrate all interval indicators into a comprehensive score for an assessment on the overall quality of GLUE intervals. This framework has three steps: (1) normalize the matrix consisting of multiple events; (2) assign a weight to each interval indicator; and (3) generate a comprehensive score by TOPSIS.

Step 1: normalization for the matrix consisting of multiple events

An increase in B (i.e. a wider band-width) implies a GLUE interval of worse quality and such indicator is called a “negative” indicator. Inversely, an increase in CR (i.e. a higher ratio of observed flows covered by GLUE interval) denotes a GLUE interval of better quality. An indicator as CR is called a “positive” indicator. Of all indicators listed in Table 2, CR is a positive indicator and the others are all negative ones. Considering the differences in the definitions of positive and negative indicators, different formulas are employed for the
normalization of positive and negative indicators. For positive indicators, there is:

\[ b_{ij} = \frac{a_{ij} - a_{j}^{\min}}{a_{j}^{\max} - a_{j}^{\min}} \quad i = 1, 2, \ldots, n; \ j = 1, 2, \ldots, 7 \]  

(7)

and, for negative indicators, there is:

\[ b_{ij} = \frac{a_{j}^{\max} - a_{ij}}{a_{j}^{\max} - a_{j}^{\min}} \quad i = 1, 2, \ldots, n; \ j = 1, 2, \ldots, 7 \]  

(8)

where \( a_{j}^{\max} = \max(a_{1j}, a_{2j}, \ldots, a_{nj}) \) and \( a_{j}^{\min} = \min(a_{1j}, a_{2j}, \ldots, a_{nj}) \). The matrix of \( B = [b_{ij}]_{n \times m} \) is called normalized event matrix, each element of which falls into the range from 0 to 1. The \( i^{th} \) row in matrix \( B \) is the \( i^{th} \) event consisting of all (normalized) indicators and the \( j^{th} \) row in matrix \( B \) consists of the \( j^{th} \) (normalized) indicators from all events.

**Step 2: weight assignment to each indicator**

The weights of indicators are obtained by CRITIC (criteria importance through inter-criteria correlation) method (Diakoulaki et al. 1995; Rostamzadeh et al. 2018). CRITIC method is able to consider both the inter- and intra-correlations of indicators. The main steps of CRITIC method are shown as follows:

1) Calculate the standard deviation of each indicator (i.e. the inter-correlations of each indicator) based on matrix \( B \):

\[ \sigma_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (b_{ij} - \bar{b}_j)^2} \quad j = 1, 2, \ldots, 7 \]  

(9)

where \( \bar{b}_j \) denotes the mean of the \( j^{th} \) indicator in matrix \( B \).

2) Compute correlation coefficient for each two indicators (i.e. the intra-correlations of indicators):

\[ \rho_{jk} = \frac{\sum_{i=1}^{n} (b_{ij} - \bar{b}_j)(b_{ik} - \bar{b}_k)}{\sqrt{\sum_{i=1}^{n} (b_{ij} - \bar{b}_j)^2 \sum_{i=1}^{n} (b_{ik} - \bar{b}_k)^2}} \quad j, k = 1, 2, \ldots, 7 \]  

(10)

3) Construct a label \( Z_j \) for each indicator by combining standard deviation with correlation coefficients:

\[ Z_j = \sigma_j \cdot \sum_{k=1}^{7} (1 - \rho_{jk}) \quad j = 1, 2, \ldots, 7 \]  

(11)

4) Lastly, obtain the weight of the \( j^{th} \) indicator by:

\[ w_j = \frac{Z_j}{\sum_{k=1}^{m} Z_k} \quad j = 1, 2, \ldots, 7 \]  

(12)
Step 3: Construction of a comprehensive score by TOPSIS

TOPSIS technique is a powerful ranking tool (Jahanshahloo et al. 2006; Kelemenis and Askounis 2010; Şengül et al. 2015). This work employs TOPSIS to yield a comprehensive score (CS) for an event consisting of multiple interval indicators. The main processes of TOPSIS are:

1) Multiply each element in matrix $B$ with CRITIC weight:

$$v_{bj} = w_j \cdot b_{ij} \quad i = 1, 2, \cdots, n; \quad j = 1, 2, \cdots, 7$$

2) Define the ideal and negative-ideal events as:

$$V_{\text{ideal}} = (V_1^-, V_2^-, \cdots, V_7^-)$$

$$V_{\text{negative-ideal}} = (V_1^+, V_2^+, \cdots, V_7^+)$$

where $V_j^+ = \max(v_{b1j}, v_{b2j}, \cdots, v_{bnj})$ and $V_j^- = \min(v_{b1j}, v_{b2j}, \cdots, v_{bnj}) (j=1, 2, \cdots, 7)$.

3) Calculate the distances from the $i^{th}$ event to the ideal and negative-ideal events by Euclidean distance as:

$$S_i^- = \sqrt{\sum_{j=1}^{7} (v_{bj} - V_j^-)^2} \quad i = 1, 2, \cdots, n$$

$$S_i^+ = \sqrt{\sum_{j=1}^{7} (v_{bj} - V_j^+)^2} \quad i = 1, 2, \cdots, n$$

4) Construct a comprehensive score (CS) based on the relative closeness between the $i^{th}$ event and the negative-ideal event:

$$CS_i = 1 - \frac{S_i^-}{S_i^+ - S_i^-} = \frac{S_i^+}{S_i^+ - S_i^-} \quad i = 1, 2, \cdots, n$$

$CS$ falls into the range from 0 to 1. An event close to ideal event (i.e. $S_i^- \to 0$) leads to $CS \to 1$ and an event close to negative-ideal event (i.e. $S_i^+ \to 0$) leads to $CS \to 0$.

5) Rank the events in matrix $B$ according to the magnitude of $CS$. A larger $CS$ implies a GLUE interval of better overall quality.

The $CS$ generated by TOPSIS quantitatively evaluates an event consisting of multiple indicators from a statistical perspective. Each event in matrix $B$ includes a range of interval indicators characterizing different geometric features of a GLUE interval. It is, therefore, statistically reasonable to regard $CS$ as a quantitative measure for the overall quality of a GLUE interval.
3. Study area, data and hydrological model

3.1 Study area and data

The source region of the Yellow River basin (SYR) is chosen as the study area of our interest (Figure 1), which is controlled by the Tangnaihai gauge with an area of 121,000 km$^2$. The SYR region is well known as the “water tower” of the Yellow River basin, supplying 38% of total runoff in the Yellow River basin (Feng et al. 2006). As is located at the Qinghai-Tibetan Plateau, the SYR region is dominated by typical cold and semi-humid climate with an annual mean precipitation of 411.0 mm. More than 75% of annual precipitation falls between May and September. Daily flows at the Tangnaihai gauge as well as daily precipitation and air temperature at 11 meteorological stations during 2000 to 2011 were collected from the Chinese Hydrological Bureau. The dataset of 2000 is used for warm-up, 2001-2010 for calibration and 2011 for validation.

![Fig.1 Map of the source region of the Yellow River basin](image)

3.2 Hydrological model

This work employs a lumped HBV model for simulating daily flows. This HBV model consists of four main modules including the snow accumulation and melt module, soil moisture accounting module, runoff response module and routing procedure module (Zhang and Lindström 1997). Please refer to Montero et al. (2016) and Bergstrom (1992) for more details of HBV model. Table 3 presents the ranges and descriptions of the parameters in HBV model.

We also pre-checked the performance of the HBV model in the SYR region before uncertainty analysis. SCE-UA method (Duan et al. 1999) was used for the optimization of HBV model parameters. Nash-Sutcliffe
coefficients between observed flows and model outputs (using optimized model parameters) are larger than 0.90 in both calibration and validation periods. Mean relative error is 4.0% in calibration period and 9.0% in validation period. This reveals that HBV model is appropriate for the estimations of daily flows in the SYR region.

**Table 3** Parameters of HBV hydrological model along with their descriptions and ranges

| Parameter   | Description                                         | Range    |
|-------------|-----------------------------------------------------|----------|
| TT (°C)     | Threshold temperature                               | 0-1      |
| CFMAX (mm°C⁻¹d⁻¹) | Degree-day factor                                    | 1.5-4    |
| SFCF        | Snowfall correction factor                          | 0.4-1    |
| CWH         | Water holding capacity                              | 0-0.2    |
| CFR         | Refreezing coefficient                              | 0-0.1    |
| FC (mm)     | Maximum soil moisture storage                       | 90-200   |
| LP          | Threshold for evaporation reduction                 | 0.7-1    |
| BETA        | Parameter that determines relative contribution to runoff from rain or snowmelt | 1-4      |
| PERC(mm/day)| Maximum percolation from the upper to the lower groundwater box | 0-1.5    |
| UZL(mm)     | Threshold parameter                                 | 0-100    |
| K₀(d⁻¹)     | Recession coefficient of upper zone                 | 0.05-0.9 |
| K₁(d⁻¹)     | Recession coefficient of lower zone                 | 0.01-0.3 |
| K₂(d⁻¹)     | Recession coefficient of deep zone                  | 0.001-0.1|
| MAXBAS (d)  | Transformation function parameter                   | 1-10     |

**4. Results and discussions**

This section attempts to investigate the potential impacts induced by Trs and likelihood measures on 90% GLUE intervals. For this purpose, 3 numerical experiments are designed by setting different combinations of Trs and likelihood measures in the processes of GLUE:

Case 1: 11 Trs (i.e. 45, 50, 55, 60, 65, 70, 75, 80, 85, 90 and 95%) with likelihood measure specified as CE;

Case 2: 11 Trs (with the same settings as Case 1) in combination with 7 likelihood measures of different factor exponents based on SE error model, i.e. CE, IRV(0.5), IRV(1), IRV(2), EL(1), EL(5) and EL(10);

Case 3: 11 Trs (with the same settings as Case 1) in combination with 9 likelihood measures of a same factor exponent based on different error models, i.e. CE, IRV(1), EL(1), NAE, IMAE, CM, LEL(1), MRE and EMRE.

The value in parenthesis (if any) denotes the factor exponent of likelihood measure. No parenthesis or value is attached behind a likelihood measure whose formula contains no factor exponent (e.g. CE, NAE, IMAE,
MRE and EMRE). In each GLUE simulation, sampling number is set as 10,000 and the prior distributions of model parameters are all assumed to be uniform.

4.1 Case 1: Impacts induced by Trs on 90% GLUE intervals

Case 1 focuses on the impacts induced by Trs on 90% GLUE intervals. In this case, likelihood measure is specified as CE for removing the impacts induced by the differences in likelihood measures. A total of 11 Trs (i.e. 45, 50, 55, 60, 65, 70, 75, 80, 85, 90 and 95%) are used for estimation of 90% GLUE intervals.

4.1.1 Impacts on the PPD distributions of model outputs

Rank observed flows in the descending order of magnitudes and define flows within the range of >80% probability as low level, 20%-80% as medium level and < 20% as high level. This section selects a low-level flow (153.9 m³/s on February 25th, 2001) and a high-level flow (747.0 m³/s on June 13th, 2001) as examples of two time points. Figure 2 shows the PPD distributions of model outputs at these two time points with 0th, 5th, 50th, 95th and 100th percentiles marked in box plots.

![Figure 2: PPD distributions and box plots of model outputs on February 25th, 2001 and June 13th, 2001 at Tr=45, 70 and 95%](image)

Note: the box plots show the 0th percentile (i.e. lower cap), 5th percentile (i.e. bottom of the box), 50th percentile (i.e. line in the box), 95th percentile (i.e. top of the box) and 100th percentile (i.e. upper cap) of model outputs together with observed flow (i.e. dashed line).

As to the low-level flow, it can be seen in Figure 2a that there is an underestimation of PPD distribution for all Trs. Most model outputs are clustered within the range from 0 to 50.0 m³/s, leading to small values of 0th, 5th and 50th percentiles in box plot as well as the failure to cover observed flow by 90% GLUE intervals at Tr=45% and 70%. Generally speaking, a likelihood measure yields a small likelihood value when a model output is far away from observed flow. An increase in Tr can reject too large or too small model outputs and, in this case, more small model outputs are rejected because most model outputs are clustered in low flow sections. This results in a rapid decrease of the probabilities in low flow sections and, therefore, flatter PPD
distributions of model outputs. Due to above reasons, 90% GLUE interval widens with the increase of $Tr$, which finally covers the observed flow at $Tr=95\%$. On the other hand, obvious distinctions can be detected between the PPD distributions of low-level and high-level flows (Figure 2a vs. Figure 2b). It can be seen in Figure 2b that PPD distributions are not extremely skewed to low flows and observed flow is successfully covered by 90% GLUE intervals at all $Tr$s. Besides, increasing $Tr$ from 45\% to 95\% causes steeper and more symmetrical PPD distributions, leading to narrower 90% GLUE intervals.

Fig. 3 90% GLUE intervals of (a) low-level (b) medium-level and (c) high-level flow sections over calibration and validation periods.
To generalize the conclusions obtained from two time points, Figure 3 presents the low-, medium- and high-level flow sections of 90% GLUE intervals at $Tr=45$, 70 and 95%, respectively. It is obvious in Figure 3 that an increase in $Tr$ generally leads to a wider GLUE interval in low-level flow sections while a narrower GLUE interval in high-level flow sections. This agrees well with the findings of single time point in former paragraphs. In Figure 3b, 90% GLUE interval narrows slightly across the whole medium-level flow sections and moves upward in the recession phases of medium-level flow sections with an increase in $Tr$. The change patterns of 90% GLUE interval in medium-level flow sections can be, to some extent, regarded as a trade-off between the change patterns of 90% GLUE intervals in low- and high-level flow sections. In addition, 90% GLUE interval in all flow level sections tend to become more symmetrical when $Tr$ increases.

As a whole, an increase in $Tr$ leads to various change patterns of 90% GLUE intervals in different flow level sections: widening trends in low-level flow sections, upward-moving trends in the recession phases of medium-level flow sections and narrowing trends in high-level flow sections. We deem that the widening or moving trend occurs when observed flow is not covered by 90% GLUE interval and narrowing trend occurs in the opposite occasion.

4.1.2 Potential causes for the variation of interval indicators with $Tr$

The black lines shown in Figures 4 and 6 are the interval indicators of 90% GLUE intervals at different $Tr$s with likelihood measure specified as CE. It can be seen that $Tr$ is related positively to CR but negatively to the other interval indicators, implying that increasing $Tr$ leads to larger coverage ratio, smaller band-width, higher symmetry degree and shorter distances from interval centers to observed flows. This also means that the geometrical features of 90% GLUE intervals are all improved with the increase of $Tr$, which is a rather rare case in practice.

The variations of interval indicators with $Tr$ can be explained with the trade-off mechanism between the widening, moving and narrowing trends of 90% GLUE interval in different flow level sections:

1) In calibration period, 54.16% of GLUE interval narrows and 45.5% widens when $Tr$ increases from 45% to 70%, while 60.01% narrows and 39.89% widens when $Tr$ increases from 70% to 95%. In validation period, 53.7% of GLUE interval narrows and 46.3% widens when $Tr$ increases from 45% to 70%, but 61.64% narrows and 38.08% widens when $Tr$ increases from 70% to 95%. This leads to the overall decreasing trends of B and RB with increasing $Tr$ in Figures 4 and 6, revealing that the variation of B and RB with $Tr$ is caused by the trade-offs between the widening and narrowing trends of 90% GLUE intervals.

2) By comparing 90% GLUE interval at $Tr=70\%$ with that at $Tr=45\%$, 118 observed flows (101 in
calibration period and 17 in validation period) are newly covered while 13 (6 and 7) are newly excluded. By comparing 90% GLUE interval at $Tr=90\%$ with that at $Tr=70\%$, 257 observed flows (204 and 53) are newly covered and 86 (77 and 9) are newly excluded. This explains the increase trend of CR with increasing $Tr$ in both Figures 4a and 6a. The variation of CR with $Tr$ can be attributed to the trade-offs between the inclusion of observed flows by widening trends and the exclusion of observed flows by narrowing trends.

3) As was found in section 4.1.1, an increase in $Tr$ generally causes higher symmetry degree and shorter distances from interval centers (i.e. the 50$^{th}$ percentile of PPD distribution) to observed flows. This accounts for the decrease trends of interval indicators related to symmetry (i.e. $S$ and $T_s$) and deviation amplitude (i.e. $D$ and $RD$) with increasing $Tr$.

It is of interest to note that most studies found negative relations between CR and $Tr$ (Chen et al. 2013; Xiong et al. 2009), which is opposite to the findings of this work. In reality, both cases do make sense from the viewpoint of the trade-offs between the widening, moving and narrowing trends of 90% GLUE interval in different flow level sections. Trade-off mechanism is the potential cause behind the variations of interval indicators with $Tr$.

4.2 Case 2: Impacts induced by the formulas of likelihood measures on 90% GLUE intervals

Case 2 attempts to address the impacts induced by the formulas of likelihood measures on 90% GLUE intervals. Likelihood measures used in this case are all selected from SE group such that remove the impacts induced by the differences in the error models of likelihood measures, namely CE, IRV(0.5), IRV(1), IRV(2), EL(1), EL(5) and EL(10). The value in parenthesis is the factor exponent of a likelihood measure.

4.2.1 Impacts on the PPD distributions of model outputs and interval indicators

Figure 5 presents the PPD distributions and box plots of model outputs on June 13$^{rd}$, 2001 (a high-level flow) obtained by above likelihood measures at $Tr=45$, 70 and 95%. It is clear in Figures 5 that the 0$^{th}$ and 100$^{th}$ percentiles of model outputs (the upper and lower boundaries of PPD distributions) are identical for all likelihood measures when $Tr$ is fixed. This proves the conclusion of theoretical analysis in section 2.1.3 that likelihood measures based on the same error model generate identical populations of accepted parameter sets. The 5$^{th}$, 50$^{th}$ and 95$^{th}$ percentiles of model outputs as well as 90% GLUE intervals are different due to the differences in the formulas of likelihood measures, which has also been theoretically demonstrated in section 2.1.3.
Fig. 4 Interval indicators of 90% GLUE intervals obtained by likelihood measures based on a same error model in calibration period
Fig. 5 PPD distributions and box plots of model outputs on June 13th, 2001 obtained by likelihood measures based on a same error model.
The interval indicators of 90% GLUE intervals over calibration and validation periods are presented in Figures 4 and 6, respectively. Visible distinctions can be detected in interval indicators and PPD distributions obtained by different likelihood measures, which are significantly reduced by an increase of $Tr$. Particularly at $Tr=95\%$, interval indicators (in Figures 4 and 6) and PPD distributions (in Figure 5f) obtained by different
likelihood measures are almost overlapped with each other. The high similarity can be attributed to the same error model of these likelihood measures. In addition, impacts induced by the factor exponent of likelihood measure are investigated in this case. Three factor exponents (i.e. 0.5, 1 and 2) are adopted by IRV and three (i.e. 1, 5 and 10) by EL. It is evident that a larger factor exponent yields sharper and more symmetrical PPD distributions in Figures 5a and 5c, thus narrower 90% GLUE intervals and shorter distances between interval centers and observed flows in Figures 5b and 5d. This can well account for the improvements of all interval indicators with increasing factor exponent in Figures 4 and 6. Of all likelihood measures, EL(10) obtains the sharpest and most symmetrical PPD distributions at all Ts. A likely reason is that EL(10) assigns very small probabilities to model outputs far away from observed flows while large probabilities to those near observed flows. It can also be seen in Figures 4 and 6 that increasing Tr can weaken the impacts induced by the factor exponents of likelihood measures on interval indicators and PPD distributions. In conclusion, above findings reveal that impacts induced by both the formulas and factor exponents of likelihood measures are controlled by Tr. Such impacts are strong at low and moderate Trs but weak at large Trs.

Slight differences in the interval indicators of 90% GLUE intervals can be detected between calibration and validation periods. Given a Tr, most interval indicators (i.e. CR, B, S, Ts, D and RD) in validation period are slightly worse than those in calibration period. The only exception is RB, possibly because only 2.1% of observed flows in validation period belong to low level (See Figure 3a). More medium- and high-level flows lead to a larger denominator in the formula of RB (Refer to Table 2) and, therefore, a lower RB.

4.2.2 Assessment on the overall quality of 90% GLUE intervals

Interval indicators describe various geometric features of 90% GLUE intervals in quantitative manners. It is, however, a difficult task to determine the best Tr for a specific likelihood measure or the best likelihood measure for a specific Tr because good qualities (e.g. high coverage ratio, small band-width, etc.) cannot be achieved simultaneously. For instance, EL(10) obtains the best CR, S and Ts at Tr=90%, but the best B, RB, D and RD at Tr=95%. In the case of Tr=90%, CE is superior to EL(10) in CR, S, Ts and RD, but inferior to EL(10) in B, RB and S.

This work assesses the overall quality of 90% GLUE intervals by integrating all interval indicators into a comprehensive score (i.e. CS) by the MADM framework introduced in section 2.4. Figure 7 shows the CSs of GLUE intervals obtained by different Trs and likelihood measures over calibration and validation periods. There is a growth in the magnitudes of CS with increasing Tr for all likelihood measures in both calibration and validation periods. Given a Tr of less than 70%, great differences can be detected in the CSs obtained by
different likelihood measures and EL(10) obtains a much larger CS than the others. Such differences can be reduced rapidly with an increase in Tr and the CSs of all likelihood measures are almost identical at Tr=95%. This reveals that the formulas of likelihood measures have few impacts on the overall quality of 90% GLUE intervals when Tr has been specified as a large value. Considering that a larger CS implies a GLUE interval of better overall quality, a large Tr is recommended for the uncertainty analysis of the daily flows simulated by HBV model in the SYR region.

![Figure 7](image_url)

**Figure 7** Comprehensive scores (CSs) of 90% GLUE intervals by likelihood measures based on a same error model

### 4.3 Case 3: Impacts induced by the error models of likelihood measures on 90% GLUE intervals

Case 3 pays attention to the impacts induced by the error models of likelihood measures on 90% GLUE intervals. The likelihood measures listed in Table 1 all are used for comparison and their factor exponents (if any) are specified as 1, i.e. CE, IRV(1), EL(1), NAE, IMAE, CM, LEL(1), MRE and EMRE.

#### 4.3.1 Impacts on the PPD distributions of model outputs and interval indicators

Figure 8 presents the PPD distributions and box plots of model outputs on June 13th, 2001 obtained by different likelihood measures at Tr=45, 70 and 95%. Figures 9 and 10 exhibit the interval indicators of 90% GLUE intervals at different Trs in calibration and validation periods. In Figures 8, 9 and 10, high similarities can be found in PPD distributions and interval indicators obtained by likelihood measures based on a same error model, but clear distinctions in those obtained by likelihood measures based on different error models. For example, both CE and IRV(1) belong to SE group and NAE belongs to AE group. PPD distributions (In Figure 8) and interval indicators (In Figures 9 and 10) are almost the same for CE and EL, whereas evidently different for CE and NAE.

This can be attributed to the formulas and characteristics of error models:

1) both SE and AE error models are based on the residuals between model outputs and observed flows.
High consistency in PPD distributions can, therefore, be detected for likelihood measures based on these two error models, i.e. CE, IRV(1), EL(1), NAE and IMAE. As SE error model is the second power form of AE error model, SE error model assigns higher probabilities to large flows and lower probabilities to low flows compared with AE error model. As a result, PPD distributions obtained by likelihood measures based on SE error model are slightly sharper and more biased to large flows than those obtained by likelihood measures based on AE error model.

Fig.8 PPD distributions and box plots of model outputs on June 13th, 2001 obtained by likelihood measures based on different error models

2) SRE error model makes use of the residuals between the square roots of model outputs and observed flows. Observed flows range from 1.4 m$^3$/s to 2739 m$^3$/s, the square roots of which range from 1.18 to 52.34. Taking square roots can reduce the impacts of large flows and assign higher probabilities to low flows.
Fig. 9 Interval indicators of 90% GLUE intervals obtained by likelihood measures based on different error models in calibration period.

3) SLE error model takes the logarithms of both model outputs and observed flows in the calculation of residuals. The logarithms of observed flows have a range from 0.33 to 7.92, which is even smaller than the square roots of observed flows. This further weakens the impacts of large flows and strengthens the impacts of low flows. Accordingly, the PPD distributions obtained by likelihood measure based on SLE error model...
(i.e. LEL) are more biased to low flows compared with those obtained by likelihood measures based on the other error models in Figure 8.

Fig. 10 Interval indicators of 90% GLUE intervals obtained by likelihood measures based on different error models in validation period

4) ARE error model divides the residuals between model outputs and observed flows by the magnitude of observed flows for removing the impacts induced by the magnitude of observed flows. By this way, ARE error model treats all flows evenly and, therefore, produces the flattest PPD distributions at all $T_r$s in Figure 8.
Based on above analysis, it can be concluded that the characteristics of PPD distributions depend largely on the error models of likelihood measures. The likelihood measures based on a same error model exhibit high similarity in obtained PPD distributions, interval indicators and 90% GLUE intervals.

Due to the differences in the formulas of likelihood measures, slight distinctions can still be detected in interval indicators obtained by likelihood measures based on a same error model. Figures 9 and 10 show that such distinctions are much more remarkable at low $T_r$s than at large $T_r$s. Considering that the error models of likelihood measures lead to similarities and the formulas of likelihood measures lead to differences, it can be concluded here that increasing $T_r$ can reduce the impacts induced by the formulas of likelihood measures whereas strengthens the impacts induced by the error models of likelihood measures. In addition, much less differences in PPD distributions and interval indicators can be found for likelihood measures within a group than those among groups. This suggests that the error model of likelihood measure exhibits more powerful impacts on 90% GLUE intervals than the formulas of likelihood measures.

4.3.2 Assessment on the overall quality of 90% GLUE intervals

Conflicts in interval indicators can be detected in Figures 9 and 10 among likelihood measures based on different error models, which are more serious than Figures 4 and 6. Figure 11 shows the CSs of 90% GLUE intervals in calibration and validation periods. It is evident in Figure 11 that increasing $T_r$ creates a larger CS for all likelihood measures, implying an improvement in the overall quality of GLUE intervals. In addition, the CSs of likelihood measures based on a same error model show high consistency and such consistency is further enhanced with the increase of $T_r$. Ranking the CSs at $T_r=95\%$ in calibration period obtains SE group $>$ AE group $>$ SRE group $>$ SLE group $>$ ARE group. This, to some extent, proves the rationality of adopting a likelihood measure based on SE error model for GLUE simulations in many works (Chen et al. 2012). There are, however, no large differences in the CSs of different likelihood measures at $T_r=95\%$ in both calibration and validation periods. The deviation between the largest and lowest CSs is less than 0.05 (expect the CS of LEL). LEL error model works poorly because it weakens the impacts induced by the large and low flows of great hydrological significances (As was mentioned, LEL error model treats all flows evenly). Based on the magnitudes of CSs, a likelihood measure based on SE error model combined with a large $T_r$ is recommended for the estimation of 90% GLUE intervals by HBV model in the SYR region.
5. Conclusions

This paper presents an investigation on the potential mechanisms behind the impacts induced by Trs and likelihood measures on 90% GLUE intervals. The SYR region is chosen as our study area and HBV model is employed for the estimation of daily flows. Seven typical interval indicators are adopted to characterize the various geometric features of 90% GLUE intervals, which are integrated into a comprehensive score by an MADM framework for an overall assessment. Major findings are summarized as follows:

1) Nine typical likelihood measures are classified into five groups according to error models, i.e. square error (SE), absolute error (AE), square root error (SRE), square logarithmic error (SLE) and absolute relative error (ARE). Likelihood measures based on a same error model are mathematically transferable, leading to an identical population of accepted parameter sets in step 3. The formulas of likelihood measures determine the PPD distributions of model outputs in step 4.

2) An increase in Tr leads to different change patterns of 90% GLUE intervals in the low-, medium- and high-level flow sections. GLUE interval widens in low-level flow sections, moves upward in the recession phases of medium-level flow sections and narrows in high-level flow sections. Trade-off mechanism among the widening, moving and narrowing trends of 90% GLUE interval is the essential reason for the variations of interval indicators with Tr. Based on the uncertainty analysis of daily flow simulation for the SYR region, it is found that increasing Tr leads to a larger coverage ratio, narrower band-width, higher symmetry degree and smaller deviation amplitude, indicating an improvement in the overall quality of 90% GLUE intervals.

3) Both the formulas and error models of likelihood measures show impacts on the PPD distributions of model outputs, interval indicators and 90% GLUE intervals. The general characteristics of PPD distributions
depend highly on the error model of likelihood measure. Much higher similarity can be detected in the PPD distributions and interval indicators obtained by likelihood measures within a group than those obtained by likelihood measures among groups. This indicates that the error model of likelihood measure has much more impacts on 90% GLUE intervals than the formula of likelihood measure. In addition, increasing $Tr$ further highlights the impacts induced by the error model of likelihood measure and weakens the impacts induced by the formula of likelihood measure. However, the impacts induced by both the formulas and error models of likelihood measures can be reduced rapidly with the increase of $Tr$.

4) Visible conflicts can be found in the interval indicators of 90% GLUE intervals obtained by different likelihood measures. This work attempts to deal with this problem by integrating all interval indicators into a single comprehensive score $CS$ by the MADM framework proposed in section 2.2.2. A larger $CS$ generally implies a GLUE interval of better overall quality. According to the magnitudes of $CS$s, a likelihood measure based on SE error model combined with a large $Tr$ is highly recommended for the estimation of 90% GLUE intervals by HBV model in the SYR region.

In conclusion, the formula of likelihood measure controls the individuality of GLUE intervals, the error model of likelihood measure controls the commonality of GLUE intervals and $Tr$ acts as a regulator between individuality and commonality. These three factors jointly result in the features of estimated GLUE intervals.

One of our future works is to testify above conclusions in other watersheds.

References:

Alazzy AA, Lü H, Zhu Y (2015) Assessing the uncertainty of the Xinanjiang rainfall-runoff model: effect of the likelihood function choice on the GLUE method. J HYDROL ENG 20(10): 04015016. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001174

Bergström S (1992) The HBV model: Its structure and applications. Swedish Meteorological and Hydrological Institute.

Beven K (2006) A manifesto for the equifinality thesis. J HYDROL 320(1-2): 18-36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Beven K, Binley A (1992) THE FUTURE OF DISTRIBUTED MODELS: MODEL CALIBRATION AND UNCERTAINTY PREDICTION. HYDROL PROCESS 6: 279-298. https://doi.org/10.1002/hyp.3360060305

Beven K, Freer J (2001) Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J HYDROL 249(1): 11-29. https://doi.org/10.1016/S0022-1694(01)00421-8

Blasone RS, Madsen H, Rosbjerg D (2008) Uncertainty assessment of integrated distributed hydrological models using GLUE with Markov chain Monte Carlo sampling. J HYDROL 353(1): 18-32. https://doi.org/10.1016/j.jhydrol.2007.12.026

Chen X, Yang T, Wang X, Xu C-Y, Yu Z (2013) Uncertainty inter-comparison of different hydrological models in simulating extreme flows. WATER RESOUR MANAG 27(5): 1393-1409. https://doi.org/10.1007/s11269-012-0244-5

Christensen S (2004) A synthetic groundwater modelling study of the accuracy of GLUE uncertainty intervals. HYDROL RES 35(1): 45-59. https://doi.org/10.2166/nh.2004.0004

Cryer SA, Applequist GE (2003) Direct treatment of uncertainty: II—Applications in pesticide runoff, leaching and spray drift
exposure modeling. ENVIRON ENG SCI 20(3): 169-181. https://doi.org/10.1089/109287503321671384

Diakoulaki D, Mavrotas G, Papayannakis I (1995) Determining objective weights in multiple criteria problems: The critic method. COMPUT OPER RES 22(7): 763-770. https://doi.org/10.1016/0305-0548(94)00059-H

Feng J, Wang T, Xie C (2006) Eco-environmental degradation in the source region of the Yellow River, Northeast Qinghai-Xizang Plateau. ENVIRON MONIT ASSESS 122(1): 125-143. https://doi.org/10.1007/s10661-005-9169-2

Freer J, Beven K, Ambroise B (1996) Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. WATER RESOUR RES 32(7): 2161-2173. https://doi.org/10.1029/95WR03723

Freni G, Mannina G, Viviani G (2008) Uncertainty in urban stormwater quality modelling: the effect of acceptability threshold in the GLUE methodology. WATER RES 42(8): 2061-2072. https://doi.org/10.1016/j.watres.2007.12.014

Freni G, Mannina G, Viviani G (2009) Uncertainty in urban stormwater quality modelling: The influence of likelihood measure formulation in the GLUE methodology. SCI TOTAL ENVIRON 408(1): 138-145. https://doi.org/10.1016/j.scitotenv.2009.09.029

He J, Dukes MD, Jones JW, Graham WD, Judge J (2009) Applying GLUE for estimating CERES-Maize genetic and soil parameters for sweet corn production. T ASABE 52(6): 1907-1921. https://doi.org/10.13031/2013.29218

He J, Jones JW, Graham WD, Dukes MD (2010) Influence of likelihood function choice for estimating crop model parameters using the generalized likelihood uncertainty estimation method. AGR SYST 103(5): 256-264. https://doi.org/10.1016/j.agsy.2010.01.006

Houska T, Multsch S, Kraft P, Frede HG, Breuer L (2014) Monte Carlo-based calibration and uncertainty analysis of a coupled plant growth and hydrological model. BIOGEO SCIENCES 11(7): 2069-2082. https://doi.org/10.5194/bg-11-2069-2014

Hwang CL, Yoon K (1981) Multiple attribute decision making methods and applications. Berlin: Springer.

Jahanshahloo GR, Lotfi FH, IzadiKhah M (2006) An algorithmic method to extend TOPSIS for decision-making problems with interval data. APPL MATH COMPUT 175(2): 1375-1384. https://doi.org/10.1016/j.amc.2005.08.048

Jensen JB (2003) Parameter and uncertainty estimation in groundwater modelling. PhD thesis, Department of Civil Engineering, Aalborg University, Series Paper No. 23.

Jin X, Xu C-Y, Zhang Q, Singh VP (2010) Parameter and modeling uncertainty simulated by GLUE and a formal Bayesian method for a conceptual hydrological model. J HYDROL 383(3): 147-155. https://doi.org/10.1016/j.jhydrol.2009.12.028

Jung Y, Merwade V (2011) Uncertainty quantification in flood inundation mapping using generalized likelihood uncertainty estimate and sensitivity analysis. J HYDROL ENG 17(4): 507-520. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000476

Kasiviswanathan KS, Sudheer KP (2016) Comparison of methods used for quantifying prediction interval in artificial neural network hydrologic models. Modeling earth systems and environment 2(1), 22. https://doi.org/10.1007/s40808-016-0079-9

Kelemenis A, Askounis D (2010) A new TOPSIS-based multi-criteria approach to personnel selection. EXPERT SYST APPL 37(7): 4999-5008. https://doi.org/10.1016/j.eswa.2009.12.013.

Lamb R, Beven K, Myrabø S (1998) Use of spatially distributed water table observations to constrain uncertainty in a rainfall–runoff model. ADV WATER RESOUR 22(4): 305-317. https://doi.org/10.1016/S0309-1708(98)00020-7

Li L, Xia J, Xu C-Y, Singh VP (2010) Evaluation of the subjective factors of the GLUE method and comparison with the formal Bayesian method in uncertainty assessment of hydrological models. J HYDROL 390: 210–221. https://doi.org/10.1016/j.jhydrol.2010.06.044

Li L, Xu C-Y, Xia J, Engeland K, Reggiani P (2011) Uncertainty estimates by Bayesian method with likelihood of AR (1) &Normal model and AR (1) &Multi-normal model in different time-scales hydrological models. J HYDROL 406: 54–65. https://doi.org/10.1016/j.jhydrol.2011.05.052

Li Z, Yang T, Huang C-S, Xu C-Y, Shao Q, Shi P, Cui T (2018) An improved approach for water quality evaluation: TOPSIS-based informative weighting and ranking (TIWR) approach. ECOL INDIC 89: 356-364. https://doi.org/10.1016/j.ecolind.2018.02.014.

Mirzaei M, Huang YF, El-Shafie A, Shatirah A (2015) Application of the generalized likelihood uncertainty estimation (GLUE) approach for assessing uncertainty in hydrological models: a review. STOCH ENV RES RISK A 29(5): 1265-1273. https://doi.org/10.1007/s00477-014-1000-6
Mo S, Lu D, Shi X, Zhang G, Ye M, Wu J (2017) A Taylor expansion-based adaptive design strategy for global surrogate modeling with applications in groundwater modeling. WATER RESOUR RES 53(12): 10802-10823.  https://doi.org/10.1002/2017WR021622

Montanari A (2005) Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall–runoff simulations. WATER RESOUR RES 41(8). https://doi.org/10.1002/2004WR003826

Montero RA, Schwanenberg D, Krahe P, Lisiak D, Sensoy A, Sorman AA, Akkol B (2016). Moving horizon estimation for assimilating H-SAF remote sensing data into the HBV hydrological model. ADV WATER RESOUR 92: 248-257. https://doi.org/10.1016/j.advwatres.2016.04.011

Olsson J, Lindström G (2008) Evaluation and calibration of operational hydrological ensemble forecasts in Sweden. J HYDROL 350(1): 14-24. https://doi.org/10.1016/j.jhydrol.2007.11.010

Romanowicz RJ, Beven K (2006) Comments on generalized likelihood uncertainty estimation. RELIAB ENG SYST SAFE 91(10): 1315-1321. https://doi.org/10.1016/j.ress.2005.11.030

Rostamzadeh RJ, Ghorabaee MK, Govindan K, Esmaeili A, Nobar HBK (2018) Evaluation of sustainable supply chain risk management using an integrated fuzzy TOPSIS-CRITIC approach. J CLEAN PROD 175: 651-669. https://doi.org/10.1016/j.jclepro.2017.12.071

Sahoo MM (2018) An Integrated Hydrologic Bayesian Multi-Model Framework for Water Quality and Nutrient Analysis (Doctoral dissertation).

Şengül Ü, Eren M, Shiraz SE, Gezder V, Şengül AB (2015) Fuzzy TOPSIS method for ranking renewable energy supply systems in Turkey. RENEW ENERG 75, 617-625. https://doi.org/10.1016/j.renene.2014.10.045

Singh A, Mishra S, Ruskauff G (2010) Model averaging techniques for quantifying conceptual model uncertainty. GROUND WATER 48(5), 701-715. https://doi.org/10.1111/j.1745-6584.2009.00642.x

Sun M, Zhang X, Huo Z, Feng S, Huang G, Mao X (2016) Uncertainty and sensitivity assessments of an agricultural–hydrological model (RZWQM2) using the GLUE method. J HYDROL 534, 19-30. https://doi.org/10.1016/j.jhydrol.2015.12.045

Viola F, Noto LV. Cannarozzo M, La Loggia G (2009) Daily streamflow prediction with uncertainty in ephemeral catchments using the GLUE methodology. PHYS CHEM EARTH 34(10): 701-706. https://doi.org/10.1016/j.pce.2009.06.006

Zhang X, Zhou J, Ye L, Zeng X, Chen Y (2015) Lower upper bound estimation method considering symmetry for construction of prediction intervals in flood forecasting. WATER RESOUR MANAG 29(15): 5505-5519. https://doi.org/10.1007/s11269-015-1131-7

Zheng Y, Keller AA (2007) Uncertainty assessment in watershed–scale water quality modeling and management: 1. Framework and application of generalized likelihood uncertainty estimation (GLUE) approach. WATER RESOUR RES 43(8). https://doi.org/10.1029/2006WR005345