Heteroscedastic and symmetric efficiency for hydrological model evaluation criteria

Chesheng Zhan, Jian Han, Lei Zou, Fubao Sun and Tiejun Wang

ABSTRACT

Evaluation criteria play a key role in assessing the performances of hydrological models. Most previous criteria are based on the standard least square method, which assumes model residuals to be homoscedastic and is, therefore, not suitable for assessing cases with heteroscedastic residuals. Here, we compared a heteroscedastic and symmetric efficiency (*HSE*) criterion with the Nash–Sutcliffe efficiency (*NSE*) and the heteroscedastic maximum-likelihood estimator (*HMLE*) by running a monthly water balance model with four parameters (i.e., the *abcd* model) in 138 basins located in the continental United States derived from the Model Parameter Estimation Experiment dataset. The results show that compared to the *NSE*, the *HSE* and *HMLE* are both more effective for stabilizing variance and producing more uniform performances with flow magnitude, and the latter is slightly more effective than the former on stabilizing the residual heteroscedasticity, with the aid of an additional parameter.

Key words | conceptual hydrological model, goodness of fit, heteroscedasticity, metric space, model performance assessment

Chesheng Zhan
Key Laboratory of Ecosystem Network Observation and Modeling,
Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences,
Beijing 100101,
China

Jian Han
Lei Zou (corresponding author)
Fubao Sun
Key Laboratory of Water Cycle and Related Land Surface Processes,
Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences,
Beijing 100101,
China
E-mail: *zoulei@igsnrr.ac.cn*

Tiejun Wang
Institute of Surface-Earth System Science,
Tianjin University,
Tianjin 300072,
China

INTRODUCTION

Model efficiency criteria quantitatively measure the closeness between simulated and observed time series of hydrological variables (Krause *et al.* 2005; Ewen 2011). Quantitative measures for model performance assessments are also necessary during model calibrations, as they are objective functions in most hydrological modeling applications (Gupta *et al.* 1998). A number of assessment criteria have been developed in the past several decades for hydrological studies (Legates & McCabe 1999; Krause *et al.* 2005; Moriasi *et al.* 2007; Ewen 2011; Pushpalatha *et al.* 2012; Bennett *et al.* 2013; Harmel *et al.* 2014), among which the Nash–Sutcliffe efficiency (*NSE*) (Nash & Sutcliffe 1970) has been mostly adopted (Legates & McCabe 1999; Ritter & Munoz-Carpena 2013). The NSE is mathematically equivalent to the standard least square (SLS) method (i.e., the Euclidean distance) for parameter calibration (Thyer *et al.* 2009) and is, thus, referred to here as an SLS-based performance measure. SLS-based criteria assume that the model residuals $\varepsilon(t)$ follow a Gaussian distribution with a zero mean (unbiasedness) and a constant variance (homoscedasticity) (Gupta *et al.* 1998; Beven 2006; Thyer *et al.* 2009; Schoups & Vrugt 2010). The model residual $\varepsilon(t)$ is defined as follows:

$$\varepsilon(t) = q_{\mathrm{sim}}(t) - q_{\mathrm{obs}}(t) \tag{1}$$

$$\varepsilon(t) \sim N(0, \sigma_0^2) \tag{2}$$

where t represents time, $q_{\mathrm{sim}}(t)$ represents the simulated flow, $q_{\mathrm{obs}}(t)$ represents the observed flow, and σ_0 represents the constant standard deviation. However, heteroscedasticity (i.e., inhomogeneous variance) has been increasingly recognized to be more common in hydrological modeling (Sorooshian & Dracup 1980; Gupta *et al.* 1998; Beven 2006; Schoups & Vrugt 2010; Pushpalatha *et al.* 2012).

doi: 10.2166/nh.2019.121

In addition to hydrological model residuals, Di Baldassarre & Montanari (2009) found that standard deviations in discharge measurements varied with observed discharge, indicating that heteroscedasticity might be a natural phenomenon existing in hydrological systems. Therefore, given that the assumption of homoscedasticity may not hold under certain conditions, the use of SLS-based criteria for evaluating model performance and calibrating model parameters might be problematic (Nourali *et al.* 2016). Moreover, some theoretical works were based on the assumptions of unbiasedness and homoscedasticity. For instance, Cheng *et al.* (2014) showed that the *NSE* was equivalent to a maximum-likelihood method with an identical Gaussian error distribution.

To resolve the aforementioned problems, it is essential to acquire quantitative knowledge of the heteroscedasticity of model residuals (Beven & Smith 2015) and develop variance-stabilizing techniques. It is a common practice to use transformations, such as logarithmic or square root transformations in hydrology, to adjust the variances on different flow levels (Oudin *et al.* 2006). The other method is the weighted least squares method, whose major problem is the weight calculation. For instance, Sorooshian & Dracup (1980) developed a heteroscedastic maximum-likelihood estimator (*HMLE*) method by assuming that the standard deviation $\sigma(t)$ in the model residuals followed a power function of the predicted flows $q_{sim}(t)$. In a later study, Sorooshian & Gupta (1983) modified $q_{sim}(t)$ to be replaced by the observed flows $q_{obs}(t)$ in the *HMLE* method (Yapo *et al.* 1998; Nourali *et al.* 2016). There have also been attempts to use a linear function to represent the heteroscedasticity of model residuals (e.g., Schoups & Vrugt 2010). The above methods can effectively stabilize the variance in model residuals; however, their formulas contain unknown parameters that must be inferred from the model and data via complicated iterations (Duan 1991; Sorooshian *et al.* 1993). Thus, the objective function cannot be fully determined before calibration. Meanwhile, the unknown parameters vary with different applications, leading to certain difficulties when comparing the performances of different case studies (e.g., multi-models or multi-basins). The primary objective of this study was to construct an efficiency criterion which was suitable for heteroscedastic cases without any additional parameters, and it can be computed based on only the observed and simulated flows.

Moreover, the choice of predicted or observed flows or their linear combination to represent the flow magnitude in the weighting function remains unknown (Sorooshian *et al.* 1993). This question might be answered by a more explicit and practical definition of efficiency criteria based on the metric space theory rather than the traditional vague definition (i.e., closeness). As pointed out by Guinot *et al.* (2011), the error component of an efficiency criterion should satisfy certain mathematical requirements (i.e., non-negativity, symmetry, and triangle inequality; (O'Searcoid 2006)), which can be given as follows:

$$\begin{cases} d(x, y) \geq 0 \text{ and } d(x, y) = 0 \Leftrightarrow x = y \\ d(x, y) = d(y, x) \\ d(x, y) \leq d(y, z) + d(z, x) \end{cases} \tag{3}$$

where $x, y, z \in \boldsymbol{X}$ and d is a so-called metric on $\boldsymbol{X}$.

The metric-based distance or similarity measures are widely used in many scientific fields, such as bioinformatics (Sippl 2008), transport geography (L'Hostis 2015), and computer science (Chen *et al.* 2009). However, the benefits of the metric-based definition of efficiency criteria have not been adequately discussed in the field of hydrology. For instance, (1) a linear combination of metrics is also a metric (see Appendix A, available with the online version of this paper) that provides a theoretical foundation for aggregated multi-objective calibrations (e.g., (Zhang *et al.* 2016)) and (2) the equivalence of metrics has been thoroughly studied with the metric space theory (O'Searcoid 2006), which can be applied to hydrological studies to distinguish equivalent efficiency criteria (e.g., the debate between Willmott *et al.* (2012) and Legates & McCabe (2013)). There are also other properties, such as convergence and continuity, which are likely helpful for efficiency criteria and need additional consideration. Due to the aforementioned advantages, heteroscedasticity and the metric-based definition were simultaneously taken into account in the design of the new proposed criterion.

The proposed criterion was compared to the *NSE* and *HMLE* in terms of the effectiveness of variance stabilization using a heteroscedasticity test (i.e., the White test) (White 1980). Further analysis of the residuals of the three objective functions was also conducted to investigate the advantages and disadvantages of the proposed criterion. The hydrological

model was run in large samples of catchment basins to arrive at more convincing and general conclusions of the criterion comparisons with a single simulation (Gupta *et al.* 2014).

METHODS

Nash–Sutcliffe efficiency

The *NSE* (Nash & Sutcliffe 1970) is the most popular criterion used to evaluate the performance of hydrological models and is defined as follows:

$$NSE = 1 - \frac{\sum_{t=1}^{N} [q_{obs}(t) - q_{sim}(t)]^2}{\sum_{t=1}^{N} [q_{obs}(t) - \overline{q_{obs}}]^2} = 1 - \left[\frac{RMSE}{\sigma_{obs}}\right]^2 \quad (4)$$

where $\overline{q_{obs}}$ represents the mean of the observed flows and t represents time. The *NSE* varies from $-\infty$ to 1 and generally falls within the range of 0–1 for calibrated simulations (Gupta *et al.* 2009). In Equation (4), the dimensional error term (i.e., the root mean square error (*RMSE*)) is normalized by dividing the standard deviation of the observed flows. Therefore, a value of $NSE = 0.8$, for example, means that the model explains 80% of the total variance in the observations (Ritter & Munoz-Carpena 2013). Specifically, $NSE = 1$ implies that the simulated flows perfectly match the observed flows, whereas $NSE = 0$ indicates that the performance of the simulated flows is equal to that of the mean observed flows (in other words, the overall results are reliable, but errors in process simulation are considerable).

The *NSE* was designed to mimic the coefficient of determination (R^2) by replacing the fitted line of R^2 to the 1:1 line for the *NSE*. The numerator of the *NSE* is the *RMSE*, and the denominator σ_{obs} is the variance in the observed flows; thus, the *NSE* represents the signal-to-noise ratio of a hydrograph. During model calibrations, since σ_{obs} is kept constant, the *NSE* is equivalent to the SLS method, which minimizes the mean square error (*MSE*) (Schaefli & Gupta 2007; Gupta *et al.* 2009; Thyer *et al.* 2009):

$$MSE = \frac{1}{N} \sum_{t=1}^{N} [q_{obs}(t) - q_{sim}(t)]^2 \quad (5)$$

SLS-based criteria, including the *NSE*, generally show their high sensitivity to peak flows in model calibrations, which can be largely attributed to the use of the squared differences in formulating these indices (Legates & McCabe 1999; Krause *et al.* 2005; Moriasi *et al.* 2007; Pushpalatha *et al.* 2012). It should also be emphasized here that both the *NSE* and SLS are prone to larger model residuals, which are more likely to occur in high flows because of heteroscedasticity (Sorooshian & Dracup 1980; Gupta *et al.* 1998; Schoups & Vrugt 2010; Pushpalatha *et al.* 2012). The characteristic results in the high sensitivity of *NSE*/SLS schemes to high flows. Meanwhile, as pointed out by Schaefli *et al.* (2005); Schaefli & Gupta (2007) and Muleta (2012), in the case of strong seasonal flows, the *NSE* overestimates the model performance, while in the case of weak seasonal flows, the *NSE* underestimates model performance. This result suggests that variance alone is not a perfect benchmark for calibrating hydrological models (Legates & McCabe 2013).

Heteroscedastic maximum-likelihood estimator

To address the issue of heteroscedasticity, Sorooshian & Dracup (1980) proposed the *HMLE*, which is calculated as follows:

$$HMLE = \frac{\frac{1}{N} \sum_{t=1}^{N} \omega(t)[q_{obs}(t) - q_{sim}(t)]^2}{\left[\prod_{t=1}^{N} \omega(t)\right]^{1/N}} \quad (6)$$

where $\omega(t)$ represents the weight depending on the flow magnitude at each time step and is computed as follows:

$$\omega(t) = [f(t)]^{2(\lambda-1)} \quad (7)$$

where $f(t)$ represents either $q_{sim}(t)$, $q_{obs}(t)$, or $\alpha q_{obs}(t) + \beta q_{sim}(t)$, with $\alpha + \beta = 1$ (Sorooshian & Dracup 1980; Sorooshian & Gupta 1983; Sorooshian *et al.* 1993; Yapo *et al.* 1996). In this study, we used $q_{obs}(t)$, as recommended by Sorooshian *et al.* (1993). As discussed previously, the unknown λ value is estimated with an iteration procedure (see Sorooshian *et al.* 1993), which gives the *HMLE* the ability to reduce heteroscedasticity.

Heteroscedastic and symmetric efficiency

Based on the metric-based definition of the efficiency criteria and previous works on the heteroscedasticity of hydrological modeling, a heteroscedasticity and symmetric

efficiency (*HSE*) is proposed here:

$$HSE = 1 - \frac{1}{N}\sum_{t=1}^{N}\left|\frac{q_{obs}(t) - q_{sim}(t)}{q_{mean}(t)}\right| \quad (8)$$

where q_{mean} represents the mean of the observed and predicted flows:

$$q_{mean}(t) = \frac{q_{obs}(t) + q_{sim}(t)}{2} \quad (9)$$

The proof for the *HSE* as a metric is provided in Appendix B (available with the online version of this paper). The *HSE* is designed in a manner similar to the concept of the Canberra distance (Lance & Williams 1966), which is used to compute the similarity of two datasets. The *HSE* is expected to reduce the heteroscedasticity of the SLS method but to a lesser degree compared to the *HMLE* method, with an additional parameter (λ). The $1/q_{mean}$ means the weights of *HSE*.

The application condition of the *HSE* is that $q_{obs}(t) \geq 0$ and $q_{sim}(t) \geq 0$, but at most only one of them could be zero at any time step t, while the application condition of the *HMLE* is $q_{obs}(t) > 0$. Thus, the applicable scope of the *HSE* is slightly broader than that of the *HMLE*. The *HSE* varies within the bounded range of -1 to 1 according to the following equation:

$$0 \leq \frac{|q_{obs}(t) - q_{sim}(t)|}{q_{obs}(t) + q_{sim}(t)} \leq 1 \quad (10)$$

$HSE = 1$ indicates a perfect match between the simulated and observed flows, and $HSE = -1$ implies simulated flows with constant zero series. In other words, there is no effective information extracted from the model simulations if $HSE = -1$. The *HSE* calculation at a certain time step is independent from other calculations, which is referred to as temporal independence. As a result of the boundedness and temporal independence, the influence of a single time step on the *HSE* is limited within $2/N$ and, therefore, the *HSE* is a robust criterion suitable for cases with outliers.

The White test for heteroscedasticity

The White test (White 1980) is a general heteroscedasticity detection method which is widely used in econometrics (Kim *et al.* 2006). Compared to the Breusch–Pagan test (Breusch & Pagan 1979), this test does not rely on the assumptions of normality or the linear form of heteroscedasticity. In this study, the White test for a single-independent variable was used to test the independence between residual variances and observed flows. The auxiliary regression of the White test is expressed as follows:

$$e^2(t) = k_0 + k_1 Q_{obs}(t) + k_2 Q_{obs}^2(t) \quad (11)$$

where $e^2(t)$ represents either squared residuals in the *NSE* method or squared weighted residuals in the *HMLE* and *HSE* methods. The greater the determination coefficient of the auxiliary equation, the more heteroscedastic $e^2(t)$ is against the observed flows. If the White statistic (NR^2) satisfies Equation (12), the null hypothesis of homoscedasticity is true; otherwise, it is rejected.

$$NR^2 <= \chi_\alpha^2(k-1) \quad (12)$$

where N represents the sample size; R^2 represents the coefficient of the determination of the auxiliary regression; α represents the significance level (set to 0.01 in this study); and k represents the degree of freedom of the auxiliary regression, with $k = 3$ (including the constant term k_0, the linear coefficient k_1, and the quadratic coefficient k_2).

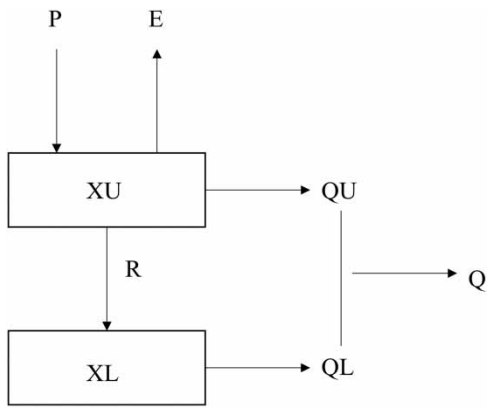

Figure 1 | Structure of the *abcd* model. Where P is monthly precipitation, E is evapotranspiration, R is groundwater recharge, QU is direct flow, QL is base flow, Q is total flow, XU is storage in the upper soil zone, and XL is storage in the lower soil zone.

Table 1 | Parameters and their ranges of the *abcd* model

Parameters	Units	Description	Ranges	Initial value
a	–	Propensity for flow to occur before the soil is completely saturated	(0,1)	0.99
b	mm	Upper soil zone water storage capacity	(0,5000)	400
c	–	Groundwater recharge coefficient	(0,1)	0.2
d	–	Groundwater discharge constant	(0,1)	0.8

Note: The initial values were determined according to the median values of parameters in the 138 basins in a rough beforehand calibration.

MODEL AND DATASET

Description of the *abcd* model

A commonly used monthly water balance model (i.e., the *abcd* model) (Thomas 1981) is selected in this study because it has been applied over a variety of hydroclimatic regions (e.g., Martinez & Gupta 2010). The model structure is illustrated in Figure 1, and the parameters with their descriptions are listed in Table 1. Detailed information on the *abcd* model can be found in Thomas (1981), and a brief description of this model is provided in Appendix C (available with the online version of this paper).

Study region and dataset

The Model Parameter Estimation Experiment (MOPEX) dataset (Duan *et al.* 2006) (available at ftp://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/) provides the daily precipitation, potential evapotranspiration, streamflow, and maximum/minimum air temperatures of 438 basins located in the continental United States. In the US MOPEX dataset, only 138 basins (Figure 2) have complete time series of monthly flow data from 1948 to 2000, with $NSE \geq 0.5$ (Moriasi *et al.* 2007), $HSE \geq 0.4$, and $|\lambda| \leq 1$, by using the *abcd* model (see the next section for details on the model calibrations). The analysis of this study focused on the ensemble results of the 138 basins. Furthermore, detailed model results of two illustrative examples are also presented to illustrate the similarity and differences among the *NSE*, *HMLE*, and *HSE*. The first catchment is the Quinebaug River Basin in Jewett City, CT (QRB;

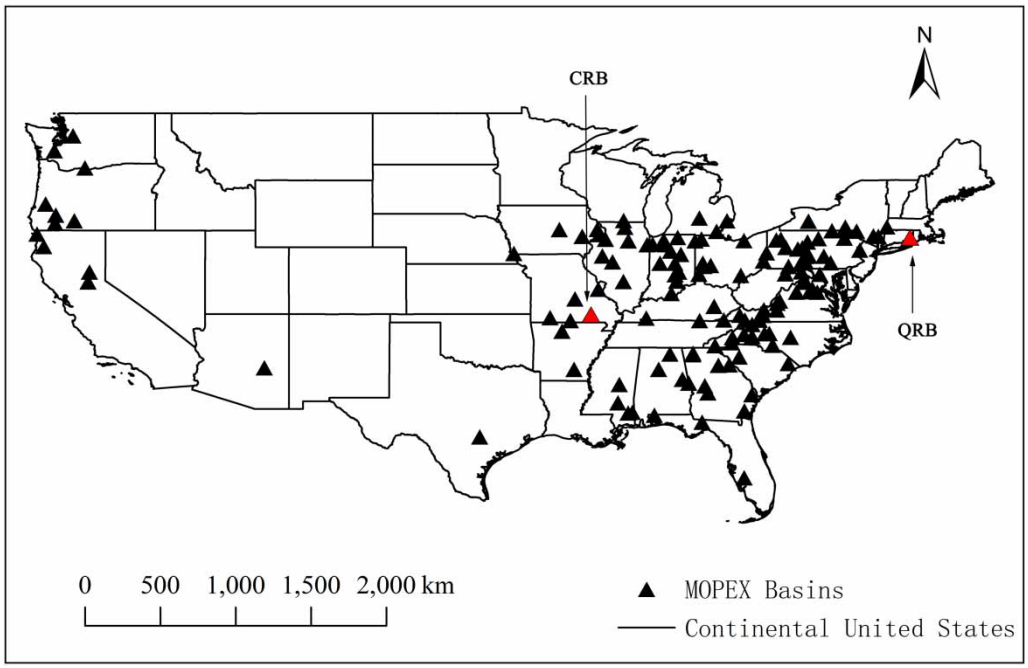

Figure 2 | Locations of the 138 MOPEX basins used in this study.

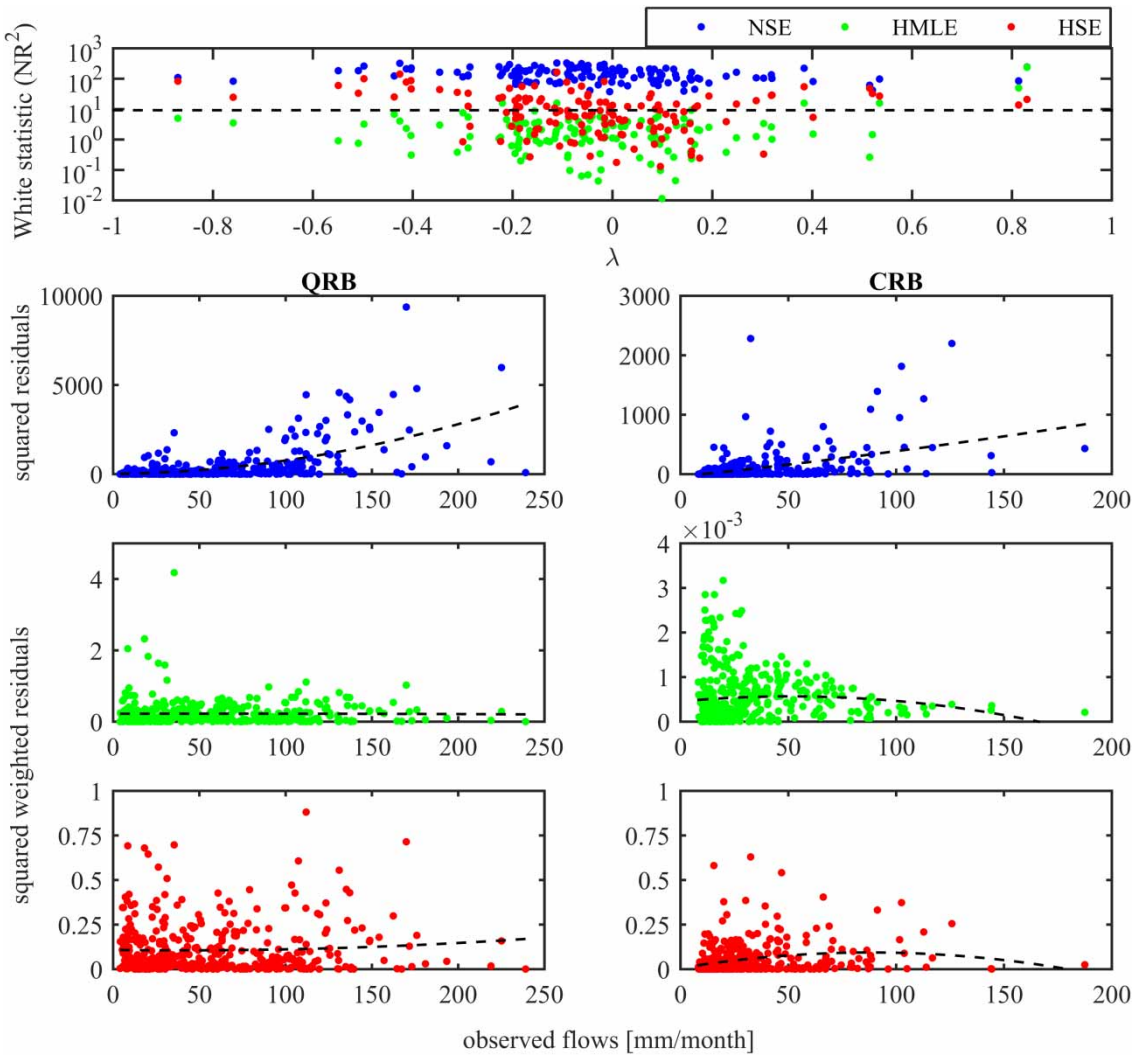

Figure 3 | White test results and scattergrams of the QRB and CRB based on NSE, HMLE, and HSE calibrations.

USGS: 01127000), and the other is the Current River Basin in Van Buren, MO (CRB; USGS: 07067000). Detailed calibration results of all 138 basins, including the simulated flows and optimal parameters with the three criteria, are provided in Supplementary Material (available online).

Calibration and validation

Martinez & Gupta (2010) compared the ranges of the four parameters of the *abcd* model derived in their study with previously reported values (Alley 1984; Vandewiele *et al.* 1992). In their studies, the parameter *a* of *abcd* model can presumably range from 0.0 to 1.0, and the values of parameter *b* <1,900 (mm/month) are more common for no-snow basins; therefore, in this study, the parameters *a*, *c*, and *d* were set to vary from 0.0 to 1.0, and parameter *b* was set to vary from 0 to 5,000 (mm/month). The shuffled complex evolution method developed at the University of Arizona (denoted as SCE-UA) was employed as the calibration algorithm (Duan, 1991). The model warm-up period was 1948–1950, the calibration period was 1951–1980, and the validation period was 1981–2000. The model parameters were, respectively, calibrated based on the criteria of *NSE*, *HMLE*, and *HSE*.

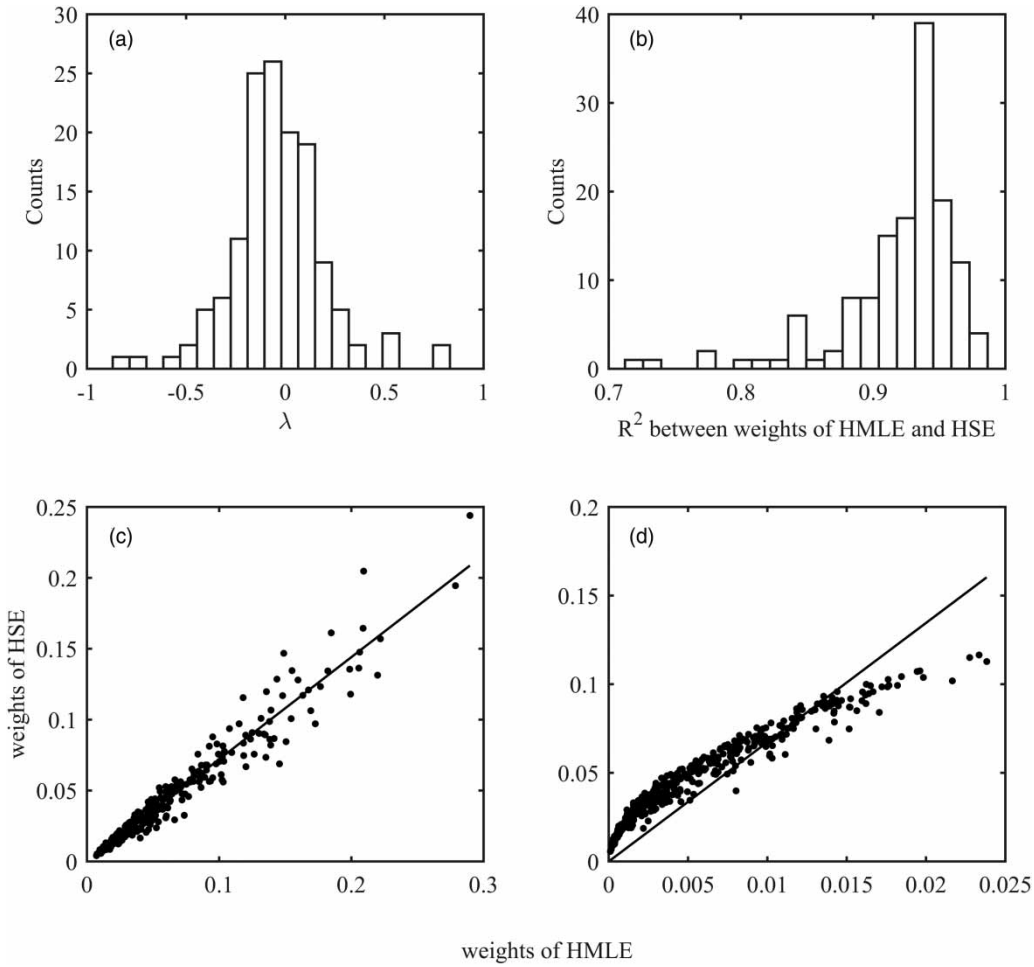

Figure 4 | Histogram of the parameter λ and relativity between HMLE and HSE weights.

RESULTS AND DISCUSSION

Effectiveness of variance stabilization

In this section, the White test was used to verify the variance-stabilizing abilities of the *NSE*, *HMLE*, and *HSE*. The results of the White test for the heteroscedasticity of the (weighted) residuals in all 138 basins and two illustrative basins are displayed in Figure 3. The *NSE* residuals in all 138 basins show significant heteroscedasticity based on the White test results (NR^2 ranges from 37.6 to 332.8, with an average of 150.0). Meanwhile, the *HMLE*- and *HSE*-weighted residuals in 95.7% (132 out of 138) and 51.5% (71 out of 138), respectively, passed those of the White test at a significance level of $\alpha = 0.01$ ($NR^2 > x_{0.01}(2) = 9.21$). Compared to the White statistics of the *NSE* residuals, the *HMLE*- and *HSE*-weighted residuals reduced the heteroscedasticity, on average, by 96.9% and 87.7%, respectively. The scattergrams of the two illustrative basins also show that the *NSE* variances were positively correlated with the flow magnitude. In comparison, the *HMLE*- and *HSE*-weighted residuals were more homoscedastic, supporting their heteroscedastic assumptions on the relationship between the variances of raw residuals and the corresponding flows. In summary, both the *HMLE* and *HSE* methods were shown to be effective for stabilizing variances for different hydrological models, while the effectiveness of the *NSE* for stabilizing variances was rather weak.

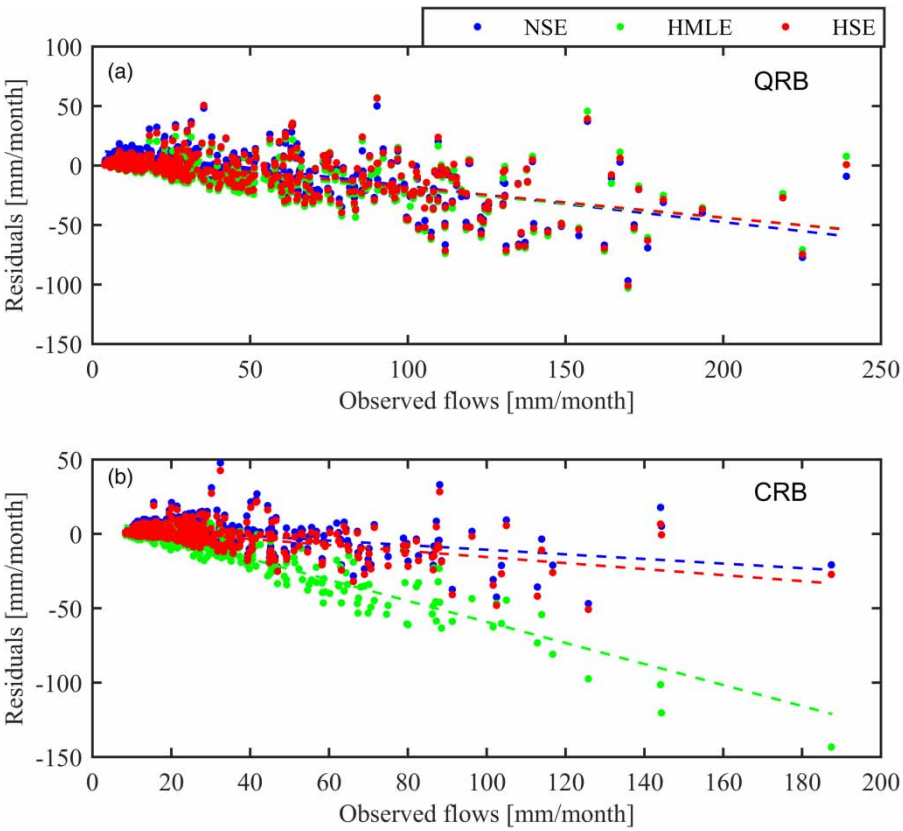

Figure 5 | Linear regression of residuals against observed flows.

Note that the parameter λ, which represents the degree of heteroscedasticity in the *HMLE* procedure, was generally close to zero, with a mean value of -0.04 and $|\lambda| \leq 0.2$ for 75% of all basins (Figure 4). Meanwhile, the weights of the *HSE* were significantly correlated with those of the *HMLE* in a positive manner ($R^2 = 0.92 \pm 0.046$), which indicates that the *HSE* method has a strong relationship with the *HMLE* method, even if the *HSE* does not use any additional parameters.

Further analysis of residuals

Ideally, the model residuals should be independent of the flow magnitudes, which usually derivates from real situations. Therefore, we regressed the model residuals against the observed flows in all 138 basins (e.g., Figure 5), and the corresponding statistical parameters (slope, intercept, and R^2) of the linear regression are shown in Figure 6. The findings are listed below:

1. The slope was mostly negative (Figure 6(b)), and the intercept was mostly positive (Figure 6(c)), which indicated that the flows were generally overestimated in low flow regimes, and vice versa.
2. The *HMLE* and *HSE* intercepts were considerably smaller than the *NSE* intercept, indicating much better performances at low flows using the *HMLE* and *HSE* than the *NSE* (Figure 7). Meanwhile, the performances of the *HSE* simulations and most of the *HMLE* simulations, where severe negative biases did not occur, were only slightly worse than the *NSE* simulations at high flows.
3. When $\lambda \geq 0$, the *NSE*, *HMLE*, and *HSE* residuals had similar R^2 values against the observed flows. However, when $\lambda < 0$, the *HMLE* residuals depended more on the observed flows than the *HSE* and *NSE* residuals, and R^2 values for the *HMLE* even reached up to 0.6–0.9, which was not expected in the model evaluation. Moreover, when $\lambda < 0$, the *HMLE* slope was more negative

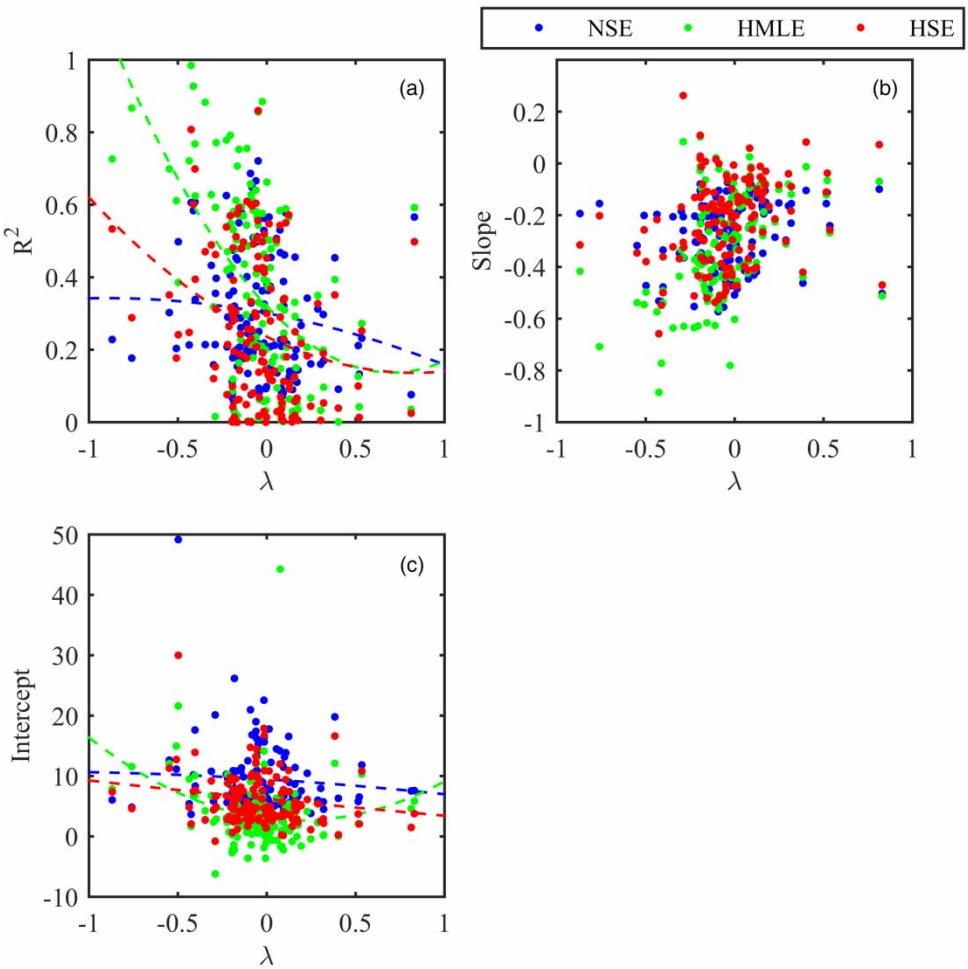

Figure 6 | Statistics parameters including the R^2, slope, and intercept of the linear regression model between residuals against the observed flow in 138 basins (λ represents the degree of heteroscedasticity in different basins).

than the *NSE* and *HSE* slopes, which led to severe negative bias at middle and high flow levels (e.g., Figures 6(b) and 8(b)).

Performance consistency between the calibration and validation periods

We compared the values of the *NSE*, *HMLE*, and *HSE* during calibration and validation. The R^2 value for the *NSE* during the calibration and validation periods was 0.61, and the R^2 values for the *HMLE* and *HSE* were 0.80 and 0.85, respectively (Figure 9). The criterion values for the *HMLE* and *HSE* (especially the *HSE*) showed great agreement between the calibration and validation periods, which is a good phenomenon in model validations. In addition, data quality, especially the rainfall input data, would influence the parameter estimation in model calibration and validation to some extent (Beven & Smith 2015). High-resolution datasets should be used to gain more precise performance.

CONCLUSIONS

Due to different weighting schemes, the *NSE* puts more emphasis on high flows, while the *HMLE* and *HSE* do not emphasize any flow levels. If modelers would like to care more about the flood peaks, the *NSE* is suggested to be used; if modelers would like to obtain more uniform

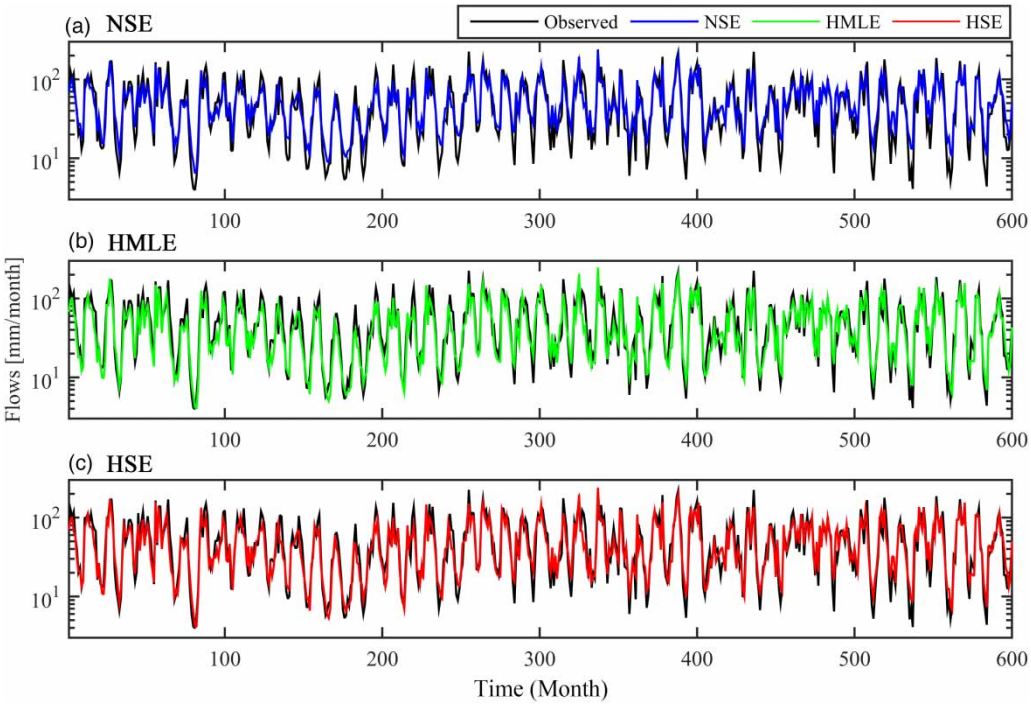

Figure 7 | Hydrographs of the Quinebaug River at Jewett City, CT (USGS: 01127000) using NSE, HMLE, and HSE.

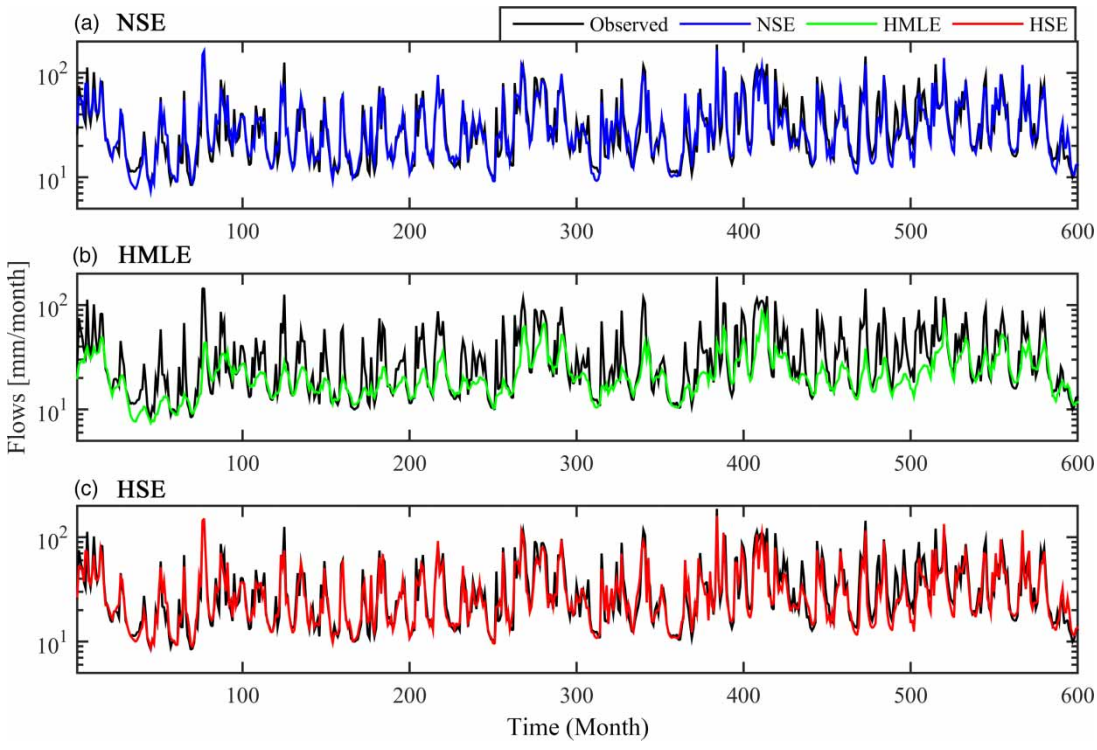

Figure 8 | Hydrographs of the Current River at Van Buren, MO (USGS: 07067000) using NSE, HMLE, and HSE.

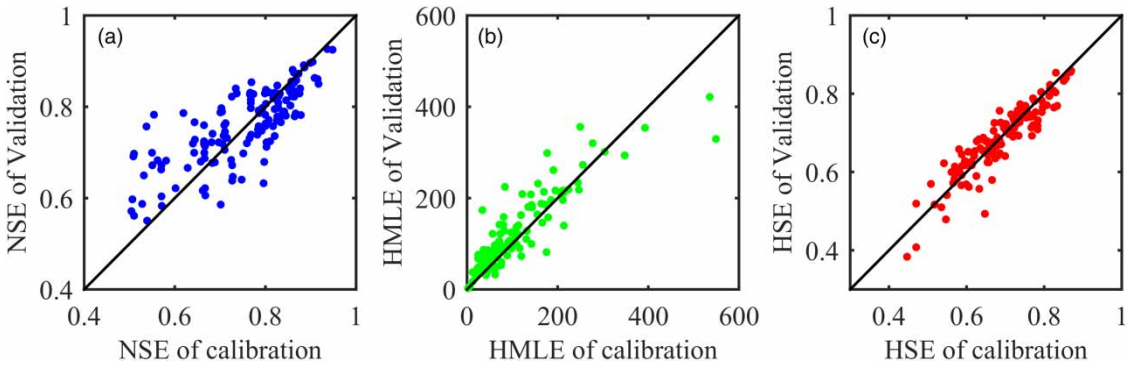

Figure 9 | Performance consistency between calibration and validation periods using NSE, HMLE, and HSE in 138 MOPEX basins.

performances along with flow magnitude, either the *HMLE* or *HSE* is suggested.

For the two heteroscedastic criteria, the *HMLE* produced slightly better homoscedastic variances than the *HSE*, with the aid of an additional parameter (λ). However, this result is not accomplished without a cost to achieve variance stabilization. When $\lambda < 0$, the *HMLE* tends to allot excessive weights for low flows and insufficient weights for middle and high flows, which might cause severe bias at middle and high flow levels. Compared to the *HMLE*-weighting approach, the *HSE*-weighting approach is slightly less effective for variance stabilization but produces more robust simulations due to the use of mean flows (q_{mean}) instead of observed flows and setting λ as zero. The robustness of the *HSE* is also embodied in two other aspects. The first aspect is that the influence of a single time step on the *HSE* is limited within $2/N$, and the second aspect is that the *HSE* during calibration and validation shows more agreement than the *NSE* and *HMLE*. Therefore, if modelers would like to stabilize the residual heteroscedasticity as much as possible, the *HMLE* is recommended for use; if the modelers would like to obtain a more robust criterion that could reduce the vast majority of residual heteroscedasticity, the *HSE* is recommended. It is suggested that future research into the proposed *HSE* and *HMLE* would be employed in different hydrological models.

ACKNOWLEDGEMENT

This study was supported by the National Key R&D Program of China (2017YFA0603702) and National Natural Science Foundation of China (no. 41571019 and no. 51779009). We are very grateful to the editors and reviewers for their valuable comments and constructive suggestions that helped us to greatly improve the manuscript.

REFERENCES

Alley, W. M. 1984 On the treatment of evapotranspiration, soil-moisture accounting, and aquifer recharge in monthly water-balance models. *Water Resour. Res.* **20** (8), 1137–1149.

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T. H., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D. & Andreassian, V. 2013 Characterising performance of environmental models. *Environ. Modell. Softw.* **40**, 1–20.

Beven, K. 2006 A manifesto for the equifinality thesis. *J. Hydrol.* **320** (1–2), 18–36.

Beven, K. & Smith, P. 2015 Concepts of information content and likelihood in parameter calibration for hydrological simulation models. *J. Hydrol. Eng.* **20** (1), A4014010, 1–15.

Breusch, T. S. & Pagan, A. R. 1979 A simple test for heteroscedasticity and random coefficient variation. *Econometrica.* **47** (5), 1287–1294.

Chen, S., Ma, B. & Zhang, K. 2009 On the similarity metric and the distance metric. *Theor. Comput. Sci.* **410** (24–25), 2365–2376.

Cheng, Q. B., Chen, X., Xu, C. Y., Reinhardt-Imjela, C. & Schulte, A. 2014 Improvement and comparison of likelihood functions for model calibration and parameter uncertainty analysis within a Markov chain Monte Carlo scheme. *J. Hydrol.* **519**, 2202–2214.

Di Baldassarre, G. & Montanari, A. 2009 Uncertainty in river discharge observations: a quantitative analysis. *Hydrol. Earth Syst. Sci.* **13** (6), 913–921.

Duan, Q. 1991 *A Global Optimization Strategy for Efficient and Effective Calibration of Hydrologic Models*. PhD Thesis, The University of Arizona, Tucson, Arizona.

Duan, Q., Schaake, J., Andreassian, V., Franks, S., Goteti, G., Gupta, H. V., Gusev, Y. M., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O. N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T. & Wood, E. F. 2006 Model parameter estimation experiment (MOPEX): an overview of science strategy and major results from the second and third workshops. *J. Hydrol.* **320** (1–2), 3–17.

Ewen, J. 2011 Hydrograph matching method for measuring model performance. *J. Hydrol.* **408** (1–2), 178–187.

Guinot, V., Cappelaere, B., Delenne, C. & Ruelland, D. 2011 Towards improved criteria for hydrological model calibration: theoretical analysis of distance- and weak form-based functions. *J. Hydrol.* **401** (1–2), 1–13.

Gupta, H. V., Sorooshian, S. & Yapo, P. O. 1998 Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. *Water Resour. Res.* **34** (4), 751–763.

Gupta, H. V., Kling, H., Yilmaz, K. K. & Martinez, G. F. 2009 Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *J. Hydrol.* **377** (1–2), 80–91.

Gupta, H. V., Perrin, C., Bloschl, G., Montanari, A., Kumar, R., Clark, M. & Andreassian, V. 2014 Large-sample hydrology: a need to balance depth with breadth. *Hydrol. Earth Syst. Sci.* **18** (2), 463–477.

Harmel, R. D., Smith, P. K., Migliaccio, K. W., Chaubey, I., Douglas-Mankin, K. R., Benham, B., Shukla, S., Munoz-Carpena, R. & Robson, B. J. 2014 Evaluating, interpreting, and communicating performance of hydrologic/water quality models considering intended use: a review and recommendations. *Environ. Modell. Softw.* **57**, 40–51.

Kim, E. H., Morse, A. & Zingales, L. 2006 What has mattered to economics since 1970. *J. Econ. Perspect.* **20** (4), 189–189.

Krause, P., Boyle, D. P. & Bäse, F. 2005 Comparison of different efficiency criteria for hydrological model assessment. *Adv. Geosci.* **5**, 89–97.

Lance, G. N. & Williams, W. T. 1966 Computer programs for hierarchical polythetic classification ('Similarity analyses'). *Comput. J.* **9** (1), 60–64.

Legates, D. R. & McCabe, G. J. 1999 Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.* **35** (1), 233–241.

Legates, D. R. & McCabe, G. J. 2013 A refined index of model performance: a rejoinder. *Int. J. Climatol.* **33** (4), 1053–1056.

L'Hostis, A. 2015 *All Geographical Distances are Optimal*, hal-01140069.

Martinez, G. F. & Gupta, H. V. 2010 Toward improved identification of hydrological models: a diagnostic evaluation of the 'abcd' monthly water balance model for the conterminous United States. *Water Resour. Res.* **46** (8), W08507, 1–21.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D. & Veith, T. L. 2007 Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* **50** (3), 885–900.

Muleta, M. K. 2012 Model performance sensitivity to objective function during automated calibrations. *J. Hydrol. Eng.* **17** (6), 756–767.

Nash, J. E. & Sutcliffe, J. V. 1970 River flow forecasting through conceptual models part I – a discussion of principles. *J. Hydrol.* **10** (3), 282–290.

Nourali, M., Ghahraman, B., Pourrezabilondi, M. & Davary, K. 2016 Effect of formal and informal likelihood functions on uncertainty assessment in a single event rainfall-runoff model. *J. Hydrol.* **540**, 549–564.

O'Searcoid, M. 2006 *Metric Spaces*. Springer Science & Business Media, London.

Oudin, L., Andreassian, V., Mathevet, T., Perrin, C. & Michel, C. 2006 Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations. *Water Resour. Res.* **42** (7), W07410, 1–10.

Pushpalatha, R., Perrin, C., Le Moine, N. & Andreassian, V. 2012 A review of efficiency criteria suitable for evaluating low-flow simulations. *J. Hydrol.* **420**, 171–182.

Ritter, A. & Munoz-Carpena, R. 2013 Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. *J. Hydrol.* **480**, 33–45.

Schaefli, B. & Gupta, H. V. 2007 Do Nash values have value? *Hydrol. Process.* **21** (15), 2075–2080.

Schaefli, B., Hingray, B., Niggli, M. & Musy, A. 2005 A conceptual glacio-hydrological model for high mountainous catchments. *Hydrol. Earth Syst. Sci.* **9** (1–2), 95–109.

Schoups, G. & Vrugt, J. A. 2010 A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resour. Res.* **46** (10), W10531, 1–17.

Sippl, M. J. 2008 On distance and similarity in fold space. *Bioinformatics* **24** (6), 872–873.

Sorooshian, S. & Dracup, J. A. 1980 Stochastic parameter-estimation procedures for hydrologic rainfall-runoff models – correlated and heteroscedastic error cases. *Water Resour. Res.* **16** (2), 430–442.

Sorooshian, S. & Gupta, V. K. 1983 Automatic calibration of conceptual rainfall-runoff models – the question of parameter observability and uniqueness. *Water Resour. Res.* **19** (1), 260–268.

Sorooshian, S., Duan, Q. & Gupta, V. K. 1993 Calibration of rainfall-runoff models: application of global optimization to the Sacramento soil moisture accounting model. *Water Resour. Res.* **29** (4), 1185–1194.

Thomas Jr., H. A. 1981 *Improved Methods for National Water Assessment, Report, Contract WR15249270*. U.S. Water Resources Council, Washington, DC.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W. & Srikanthan, S. 2009 Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: a case study using Bayesian total error analysis. *Water Resour. Res.* **45** (12), W00B14, 1–22.

Vandewiele, G. L., Xu, C. Y. & Larwin, N. 1992 Methodology and comparative-study of monthly water-balance models in Belgium, China and Burma. *J. Hydrol.* **134** (1–4), 315–347.

White, H. 1980 A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48** (4), 817–838.

Willmott, C. J., Robeson, S. M. & Matsuura, K. 2012 A refined index of model performance. *Int. J. Climatol.* **32** (13), 2088–2094.

Yapo, P. O., Gupta, H. V. & Sorooshian, S. 1996 Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data. *J. Hydrol.* **181** (1–4), 23–48.

Yapo, P. O., Gupta, H. V. & Sorooshian, S. 1998 Multi-objective global optimization for hydrologic models. *J. Hydrol.* **204** (1–4), 83–97.

Zhang, Y. Y., Shao, Q. X., Zhang, S. F., Zhai, X. Y. & She, D. X. 2016 Multi-metric calibration of hydrological model to capture overall flow regimes. *J. Hydrol.* **539**, 525–538.

First received 11 February 2019; accepted in revised form 21 May 2019. Available online 20 June 2019