Pitfalls and a feasible solution for using KGE as an informal likelihood function in MCMC methods: DREAM(ZS) as an example

Yan Liu[1], Jaime Fernández-Ortega[2], Matías Mudarra[2], Andreas Hartmann[1,3]

[1]Chair of Hydrological Modeling and Water Resources, University of Freiburg, 79098 Freiburg, Germany
5 [2]Department of Geology and Centre of Hydrogeology, University of Málaga (CEHIUMA), 29071 Málaga, Spain
[3]Department of Civil Engineering, University of Bristol, Bristol, UK

Correspondence to: Yan Liu (yan.liu@hydmod.uni-freiburg.de)

Abstract. The Kling-Gupta Efficiency (KGE) is a widely used performance measure because of its advantages in orthogonally considering bias, correlation and variability. However, in most Markov chain Monte Carlo (MCMC)
10 algorithms, error-based formal likelihood functions are commonly applied. Due to its statistically informal characteristics, using the original KGE in MCMC methods leads to problems in posterior density ratios due to negative KGE values and high proposal acceptance rates resulting in less identifiable parameters. In this study we propose adapting the original KGE using a gamma distribution to solve these problems and to apply KGE as an informal likelihood function in the DiffeRential Evolution Adaptive Metropolis DREAM(ZS), which is an advanced MCMC algorithm. We compare our results with the
15 formal likelihood function to show whether our approach is robust and plausible to explore posterior distributions of model parameters and to reproduce the discharge behaviors. For that, we set three case studies that contain different uncertainties. Our results show that model parameters cannot be identified and the uncertainty of discharge simulations is large when directly using the original KGE. Our approach finds similar posterior distributions of model parameters compared to the formal likelihood function. Even though the acceptance rate of the adapted KGE is lower than the formal likelihood function
20 for some systems, the convergence rate (efficiency) is similar between the two approaches for the calibration of real hydrological systems showing generally acceptable performances. We also show that both the adapted KGE and the formal likelihood function provide low performances for low flows, with the larger overestimations obtained from using the formal likelihood function. Furthermore, the adapted KGE approach behaves closely to the formal likelihood function in terms of the correlation between simulations and observations. Thus, our study provides a feasible way to use KGE as an informal
25 likelihood in the MCMC algorithm and provides possibilities to combine multiple data for better and more realistic model calibrations.

1 Introduction

Markov chain Monte Carlo (MCMC) techniques are extremely useful in uncertainty assessments and parameter estimations of hydrological models (Smith & Marshall, 2008). Among those MCMC methods, Vrugt et al. (2008, 2009) developed a

30   DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm, which has found numerous applications in various fields (Vrugt, 2016). It is an adaptation of the SCEM-UA algorithm (Vrugt et al., 2003) that can efficiently estimate the posterior probability distribution of model parameters in the presence of high-dimensional and complex response surfaces with multiple local optima.

The formal likelihood function, e.g. mean squared error (MSE) or root mean squared error (RMSE), obtained from first-

35   order statistical principles based on error series derived from simulations and observations, is commonly used in the DREAM algorithm. The formal likelihood function strongly relies on error assumptions, which can highly influence the shape of parameter posterior distributions (Beven et al., 2008). The informal likelihood functions such as Nash-Sutcliffe efficiency (NSE) and the alternative Kling-Gupta Efficiency (KGE) are often used in hydrological studies to indicate the general performance of model simulations (Gupta et al., 2009). These metrics represent an important measure of model

40   performance, so-called goodness-of-fit (Pool et al., 2018). These likelihood functions are not directly derived from stochastic error series, but can be easily used to combine different types of data.

There are studies that discussed how to adjust the calculation of NSE in order to overcome the problems using NSE in MCMC methods. For example, McMillan & Clark (2009) introduced a constant K into the adapted NSE calculation. By adjusting this constant, they can mimic the weight to small improvements in NSE. They even found that the informal

45   likelihoods can provide a more complete exploration of the behavioral regions of the response space and hence more accurate estimation of total uncertainty (McMillan & Clark, 2009). Freer et al. (1996) introduced a parameter N as an exponent. They argued that higher N values have the effect of accentuating the weight given to the better simulations.

However, how to properly use KGE in the MCMC methods has not been studied. Directly using KGE in MCMC methods, e.g. DREAM algorithm, may raise difficulties such as incorrect posterior ratios due to negative KGE values, and nonlinearity

50   between model performance and KGE values. These difficulties essentially affect chain evolutions such as the acceptance rate, indicating how easy a proposal is accepted, and the convergence rate, denoting how fast a chain converges to a stationary distribution. As a consequence, the informal character of KGE and its use in MCMC methods influences the exploration of posterior parameter distribution and model uncertainty, such as density of identifiable parameters. Studies showed that using informal likelihood functions in generalized likelihood uncertainty estimation (GLUE) may lead to

55   unsatisfied posterior distributions of model parameters (Mantovan & Todini, 2006; Stedinger et al., 2008). Using NSE as the likelihood function cannot account for the increasing information for parameter estimations such that little information will be extracted from the used observations (Mantovan & Todini, 2006). Therefore, to feasibly use KGE in MCMC methods requires solving problems in drawing better proposals to avoid a very flat posterior distribution, to account for the influence of observation size (the amount of information included in calibration) on parameter estimations and to achieve reasonable

60   acceptance and convergence rate.

In this study we propose adapting the gamma distribution and KGE to find a feasible solution for properly using KGE as an informal likelihood function in DREAM$_{(ZS)}$. We test the robustness of this approach with three case studies representing known and unknown systems with varying amount of observations using two hydrological models (a lumped and a semi-distributed). The aim of our study is to attempt to form the probability calculation based on KGE in a pseudo formal way.

65 The derivation of probability based on KGE in a theoretically statistical manner is beyond the scope of this study and will need future work. We will compare the performance between the original KGE, the formal likelihood function RMSE and the adapted KGE, but do not aim to show whether the adapted KGE approach outperforms the formal likelihood function or not. Instead, we show whether the adapted KGE is robust and plausible to explore posterior distributions of model parameters and to reproduce the discharge behaviors as similar as the formal likelihood function. Thus, we will compare

70 performance (1) regarding acceptance, convergence, and uncertainty between the adapted KGE, the original KGE and a formal likelihood function, and (2) regarding discharge simulations in terms of general performance, variability, bias and correlation for total, low and high flows. These comparisons allow us to provide recommendations for more reliable applications of KGE in MCMC methods in different research areas.

2 Material and Methods

75 ### 2.1 Kling-Gupta Efficiency

Kling-Gupta Efficiency (KGE) takes account of variability (α), non-scaled bias (β) and correlation (r) by computing the Euclidian distance (*ED*) of the three components from the ideal point, which avoids the underestimation of variability and enables the comparison of the bias term between catchments (Gupta et al., 2009).

$$KGE = 1 - ED \tag{1}$$

80 $$ED = \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{2}$$

with $\alpha = \frac{\sigma_s}{\sigma_o}$ and $\beta = \frac{\mu_s}{\mu_o}$

where (μ_s, σ_s) and (μ_o, σ_o) are the mean and standard deviation of simulations and observations, respectively. KGE ranges from $-\infty$ to 1 with the optimal value at unity. A value larger than -0.41 indicates that a model improves upon using the mean (Knoben et al., 2019).

85 ### 2.2 Adapting KGE in DREAM$_{(ZS)}$

2.2.1 Basics of DREAM$_{(ZS)}$

DREAM$_{(ZS)}$ is one type of DREAM algorithms, which uses sampling from an archive of past states to generate candidate points in each individual chain. It automatically tunes the scale and orientation of the proposal distribution towards the target

distribution and maintains a detailed balance and ergodicity (Vrugt et al., 2008, 2009). We take DREAM$_{(ZS)}$ as an example

90   and investigate the appropriate way to use KGE within DREAM$_{(ZS)}$. Since our goal is to adapt KGE as an informal

likelihood function, we focus on the Metropolis probability, $p_{acc}(X^i \rightarrow X_p^i)$, which calculates the probability to accept a

proposal.

$$p_{acc}(X^i \rightarrow X_p^i) = min[1, p(X_p^i)/p(X^i)] \tag{3}$$

where $p(X_p^i)$ and $p(X^i)$ denote the probability density of the proposal and the present location of the i-th chain. If

95   $p_{acc}(X^i \rightarrow X_p^i)$ is larger than a random value drawn from the uniform distribution $U(0, 1)$, the proposal will be accepted,

otherwise the chain remains in the present location. After chain evolutions, the Gelman-Rubin $\hat{R}_J$ convergence diagnostic is

computed for each parameter $j=\{1, \ldots, d\}$. If $\hat{R}_J \leq 1.2$, the convergence can be declared.

$$\hat{R}_J = \sqrt{\frac{N+1}{N} \cdot \frac{\hat{\sigma}_+^{2(j)}}{W_j} - \frac{T-2}{N \cdot T}} \tag{4}$$

where N and T signify, respectively, the number of chains and the number of samples in each chain. W_j is the within-chain

100 variance, and $\hat{\sigma}_+^{2(j)}$ is an estimate of the variance of the j-th parameter of the target distribution.

2.2.2 Deriving pseudo probability density based on KGE

Based on above-mentioned basics of DREAM$_{(ZS)}$, using KGE as a likelihood function becomes the proper calculation of

pseudo probability density $p(X_p^i)$. From the straightforward and easily applied way, one would directly use KGE as $p(X_p^i)$

by setting the negative KGE values to zeros. However, it results in two problems (Fig. 1):

105 • Problem 1: the probability density $p(X_p^i)$ needs to be positive in order to get a positive ratio $p(X_p^i)/p(X^i)$ to

determine the right orientation to accept a proposal. However, KGE ranges from -∞ to 1. Setting negative KGE

values to zero can work, but then we lose the orientation of proposals with negative KGE values. Thus, it reduces

the efficiency of chain evolution.

• Problem 2: the model performance does not linearly increase with the linear increase of KGE. Therefore, directly

110   using positive KGE as the pseudo probability density $p(X_p^i)$ will lead to a high possibility to accept poor proposals.

For instance, the probability is 0.75 (0.6/0.8) to accept a proposal with KGE=0.6 under the present sample with

KGE=0.8. However, the performance of a simulation with KGE=0.8 is much better compared to one with

KGE=0.6.

We propose solving these two problems with adapting the gamma distribution and KGE to derive a proper probability

115 density (Fig. 1). The gamma distribution has two parameters, the shape parameter (k) and the scale parameter (θ), and is

evaluated for variables with positive values. When the shape parameter $k=1$, the probability distribution is one-side with non-linear decreasing probability for increasing variable values. Therefore, we can use 1-KGE (ranges from 0 to ∞) as the variable for gamma distribution and get higher probability for larger KGE values. When choosing the scale parameter $\theta=0.5$, the increasing rate of probability becomes faster when KGE>0.5, especially when KGE>0.7 the probability increases much

120 faster. This helps chain evolution to find proposals which leads to high model performance. The non-linear increase of probability with increasing KGE is shown as the red line in the top right box of Fig. 1. The log-likelihood is used in DREAM$_{(ZS)}$, thus we derive the pseudo log-likelihood (LogL) using gamma distribution and KGE. To include the influence of observation size on the parameter estimations, analogy to the formal likelihood function we take account of the number of observations (the term n/2 in the formal log-likelihood) in deriving the pseudo log-likelihood. Finally, we have the pseudo

125 log-likelihood using KGE and gamma distribution as:

$$\text{LogL} = \frac{n}{2} log\big(f(1 - KGE|X; k, \theta)\big), \text{ with } k=1 \text{ and } \theta=0.5 \tag{5}$$

where, $f()$ is the gamma probability density function; k and θ are the shape and scale parameters, respectively. n is the number of observations. X represents the parameter vector of the calibrated model. The purpose of this approach is to provide a feasible way to incorporate KGE as the informal likelihood function for MCMC methods, in our case DREAM$_{(ZS)}$.

130 Another goal is to achieve similar performance as using the formal likelihood function such as RMSE so that we can compare the model simulations and predictions between formal and informal (our approach, KGE) likelihood functions.
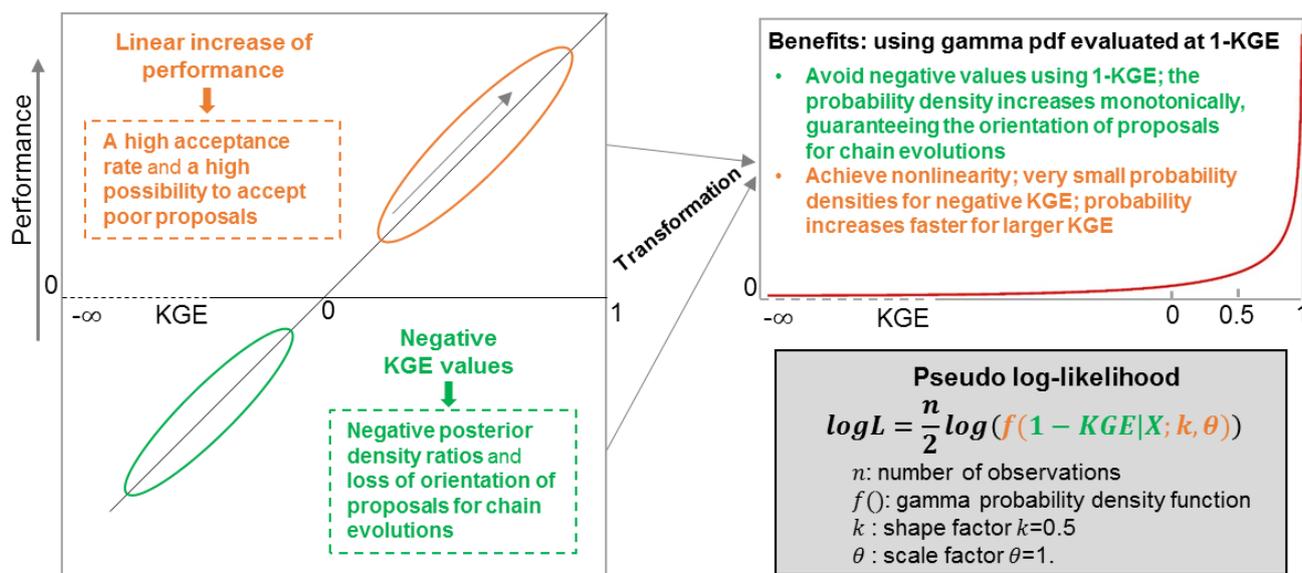
Figure 1 Concept of adapting KGE in DREAM$_{(ZS)}$. It shows the problems of using the original KGE as the likelihood function in DREAM$_{(ZS)}$ and how to adapt gamma distribution and KGE to get a proper informal likelihood.

Hydrology and
Earth System
Sciences
Discussions

2.3 Case studies

To test the robustness of our new approach, we define three case studies: (1) a single-optimum calibration, where true model parameters are known by the virtual experiment and uncertainties in model structures and input data are absent; (2) calibrations and evaluations with a long observation time-series using a rainfall-runoff model, which allows comparing the performance between three approaches by varying the amount of data in calibration. The parameterization of the system is unknown and there are uncertainties in model structure, input data and observations; and (3) a model calibration for a karst system, which is a more complex system with a large subsurface heterogeneity and processes for fast recharge and groundwater discharge from conduit networks. The observation period is short and uncertainties exist in model structure, input and observation data, and model parameter estimations.

2.3.1 Case study 1: virtual experiment

We generate a virtual experiment using a rainfall-runoff model (the HBV model). We obtain the forcing data, daily mean temperature and daily precipitation (2001–2008), from the German site in Liu et al. (2021). The HBV model represents typical catchment rainfall-runoff processes considering one soil water storage and two groundwater storages (Lindström et al., 1997). In this virtual experiment, we use the model version without snow processes, which contains nine parameters. Since our goal is not the model itself but the calibration of model parameters, we only provide descriptions of model parameters (Table 1). For the model structure and equations, refer to Liu et al. (2021).

Table 1 Names, description, ranges and virtual true values of the HBV model parameters for the virtual experiment

| Parameter	Description	Unit	Parameter ranges		True value
			lower	upper	
BETA	Shape coefficient of recharge function	[-]	1	6	4.5
FC	Maximum water storage in the unsaturated-zone store	[mm]	50	700	600
K0	Additional recession coefficient of upper groundwater store	[d^{-1}]	0.05	0.99	0.5
K1	Recession coefficient of upper groundwater store	[d^{-1}]	0.01	0.8	0.25
K2	Recession coefficient of lower groundwater store	[d^{-1}]	0.001	0.15	0.07
LP	Soil moisture value above which actual evaporation reaches potential evaporation	[-]	0.3	1	0.55
PERC	Maximum percolation to lower zone	[mmd^{-1}]	0	6	3
UZL	Threshold parameter for extra outflow from upper zone	[mm]	0	100	60
MAXBAS	Length of equilateral triangular weighting function	[d]	1	3	2

note: *K0, UZL and MAXBAS* (shaded ones) are insensitive parameters in this case study, thus are fixed to the true values

Hydrology and
Earth System
Sciences
Discussions

Open Access

EGU

We set the catchment area of 100 km^2 and run the model with "true" parameters for 2001–2008 to obtain the virtual discharge. To account for observation errors, we added random errors (-0.1σ – 0.1σ) to the virtual discharge time series to form the final virtual observations. We performed a local parameter sensitivity analysis before model calibrations to find insensitive parameters (*K0*, *UZL* and *MAXBAS* in Table 1, changing these parameters only affects model performance KGE by 0.001). Therefore, we fixed the three parameters in calibration, resulting in six parameters to be calibrated.

2.3.2 Case study 2: long observations for a rainfall-runoff modeling

In order to test the capability of our approach for a real system with uncertainties in forcing, observations, model structure and model parameters, we select a catchment from the CAMELS-US dataset (Newman et al., 2014, 2015) and simulate the rainfall-runoff processes with the HBV model as case study 1. We have the following criteria to select this catchment: 1) catchment area is between 100 and 500 km^2 to avoid the influence of channel routing; 2) the snow fraction is zero to avoid the snow processes to reduce the HBV model parameters and the number of parameters remain nine (table 1); and 3) the carbonate rock fraction is zero since we will test a karst catchment as a separate case study. We then choose the first catchment (smallest number in catchment ID) that fulfills our criteria, the catchment 02246000 (gauging station at North Fork black creek NR Middleburg, FLA, USA). It has a catchment area of 451 km^2 with the mean annual precipitation of 1352 mm. The main land cover is savannas, accounting for 77% of the total area. Details of the catchment properties can be found in CAMELS-US dataset (Newman et al., 2014, 2015).
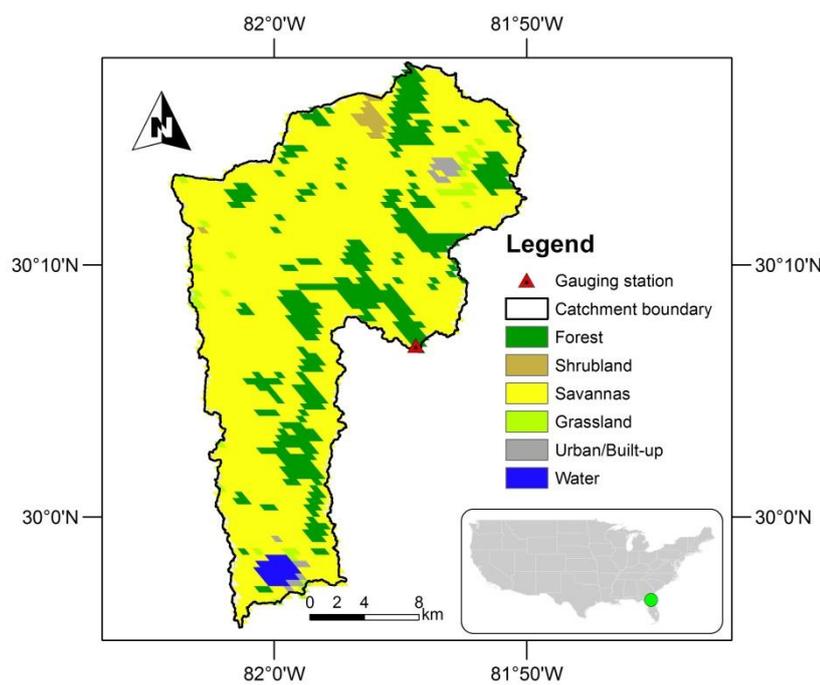

Figure 2 Study area of the catchment with the gauging station at North Fork black creek NR Middleburg, FLA, USA for the case study 2.

2.3.3 Case study 3: short observations for a heterogeneous karst system

In order to test the capability of our new approach in another type of hydrological system, the karst system that has conduit systems resulting in fast recharge and discharge, we set case study 3. It has uncertainties from the forcing data, the model

175 structure and discharge observation errors. The study site is a karst system (13.85 km²) located in Southern Spain (the study site in Hartmann et al., 2014). In this case, the VarKarst model is used for the spring discharge simulations since it was successfully applied to this site before (Hartmann et al., 2014). The VarKarst model is a semi-distributed hydrological model, which considers subsurface heterogeneity, soil and epikarst storage dynamics and groundwater hydrodynamics. These processes are represented by seven parameters (Table 2). Details of the VarKarst model can be found in Hartmann et

180 al. (2013).
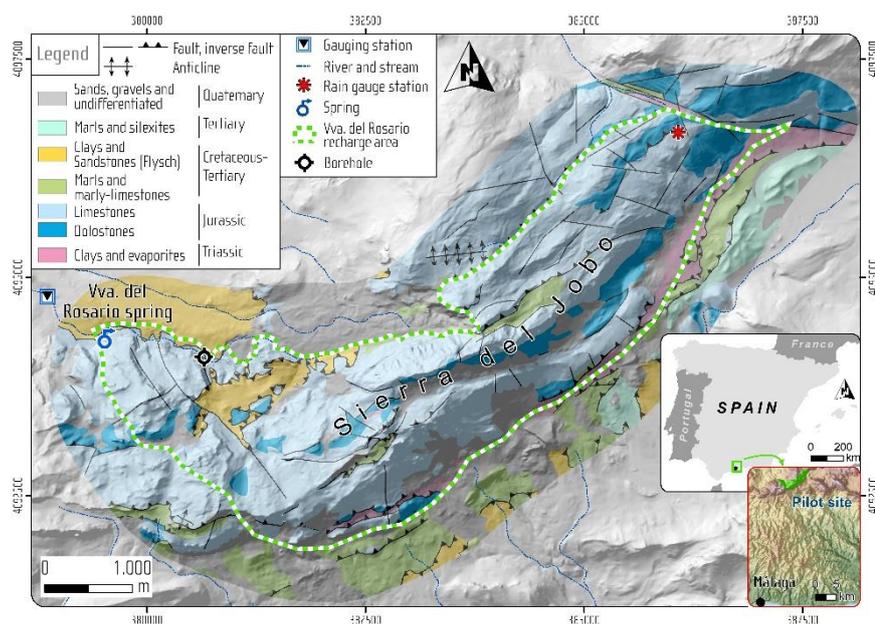
Figure 3 Study area of the Rosario spring for the case study 3. This map is an updated version of the map in Hartmann et al. (2014)

Table 2 Names, descriptions and ranges of the VarKarst model parameters

Parameter	Description	Unit	Parameter ranges	
			lower	upper
$V_{mean,S}$	Mean soil storage capacity	[mm]	0	500
$V_{mean,E}$	Mean epikarst storage capacity	[mm]	0	500
a_{SE}	Soil/epikarst depth variability constant	[-]	0.1	2
$K_{mean,E}$	Epikarst mean storage coefficient	[d]	1	50
a_{fsep}	Recharge separation variability constant	[-]	0.1	2
K_C	Conduit storage coefficient	[d]	1	20
a_{GW}	Groundwater variability constant	[-]	0.1	2

2.3.4 Calibration and evaluation

For calibration using DREAM$_{(ZS)}$, we define three strategies: (1) using the original KGE as the likelihood function – here negative KGE is set to zero to avoid negative posterior density ratios; (2) using the traditional formal likelihood RMSE; and (3) using our new approach adapting gamma distribution and KGE to derive the pseudo log-likelihood. We use three parallel Markov chains (default setting in DREAM$_{(ZS)}$) and 20,000 realizations for all three case studies. The last 25% of realizations are used to approximate the posterior distributions and the corresponding parameter sets are used for the discharge simulations in the evaluation period.

For case study 1, since the true model parameters and the initial conditions are known, no warm-up is needed. We use 1, 2, 3, 5, and 8 years of observations to perform calibrations in order to investigate how the three approaches differ for the exploration of true model parameters with varying amount of data included in calibration. For case study 2, the true model parameters are unknown. We use 25 hydrological years (10.1.1980−9.30.2005) to perform the calibration and evaluation. We choose the Daymet forcing data to drive our hydrological model since this meteorological data has potential evapotranspiration (PET) estimates and was used to calculate the catchment climatic properties (Addor et al., 2017). We use the first 5 years as the warm-up period, then the following 10 years for calibration and the last 10 years for evaluation. Similar as case study 1, we test the performance of three approaches using 1, 3, 5, 8 and 10 years of observations for calibrations, respectively to check the capability of the three approaches for the calibration of a real system with varying amount of available observations. For case study 3, the true model parameters are unknown too. We use three hydrological years (10.1.2003−9.30.2006) for warm-up of the simulations. Since there are only three hydrological years (10.1.2006−9.30.2009) of observations, we use the first half as calibration and the second half as evaluation to test the three approaches for calibration of a heterogeneous hydrological system with short available observations.

Hydrology and
Earth System
Sciences
Discussions

Open Access

EGU

205 The model performance for calibration and evaluation is examined using KGE and its three components representing variability (α), bias (β) and correlation (r). The calculations of KGE, α, β and r refer to equations 1 and 2. We evaluate the model performance using the four metrics under three flow conditions: total flow, low flow (smaller than 10th percentile of observed discharge) and high flow (larger than 90th percentile of observed discharge) for the three approaches.

3 Results

3.1 Case study 1: posterior parameter exploration

When using the original KGE (set negative KGE values to zeros) as the likelihood function, the posterior parameter range is slightly reduced for all sensitive parameters compared to the prior uniform distribution. However, the density around the true values of model parameters are still very flat, indicating true model parameters are barely identified (Fig. 4). When applying the adapted KGE (adapting the gamma distribution and KGE to derive probability density), the posterior parameter range is

215 much more reduced and the reduced range is more or less centered at the true values no matter how many observations are included in calibrations (varying information from low to high for calibrations) as shown in Fig. 4. When adding more observations to gain more information in calibrations (comparing calibrations with 1-year and 8-year data, Fig. 4), we can see that the parameter range is even narrower and the density around the true parameter values become higher. The posterior distributions of all model parameters show similar magnitudes and shapes between the adapted KGE and the formal

220 likelihood function RMSE for calibrations with different amount of observations. It suggests that our adapted KGE approach performs similarly as using the traditional formal likelihood function in terms of exploring parameter posterior distributions.
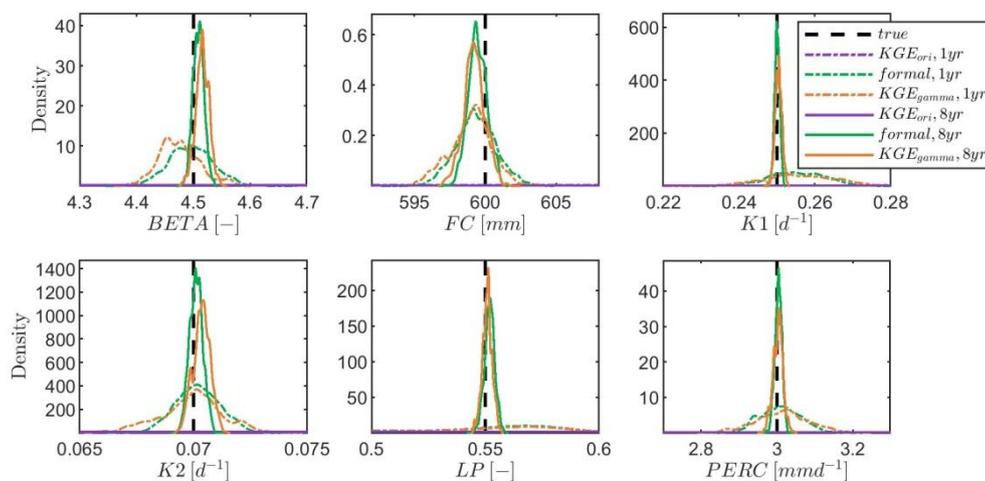

Figure 4 Posterior distributions of sensitive model parameters for the virtual experiment. The black dashed vertical line (true) denotes the true model parameter values. KGE_{ori} indicates using the original KGE (set negative KGE to zero) as the likelihood function, while
225 KGE_{gamma} represents our new approach using gamma distribution and KGE to derive the probability density, and the case *formal* means using the traditional formal likelihood function (RMSE). The subscripts *1yr* and *8yr* denote calibrations with 1-year and 8-year

observations, respectively. Note that the densities of using KGE_{ori} with 1-year and 8-year observations are both low such that they are barely distinguished and are close to the x axis.

As expected, using the original KGE, we have a very high acceptance rate (ca. 60-80%, Fig. 5a), leading to a very fast
230   convergence (Fig. 5b) for calibrations with short to long observation records. This results in a large uncertainty bound in the discharge simulations (Fig. 5c), and the uncertainty of peak discharges is particularly large. With the adapted KGE, we see that the acceptance rate becomes smaller and the convergence gets slower. This can be explained by introducing the nonlinearity of the adapted KGE: probability densities for large and small KGE values are more distinct compared to the original KGE. Fig. 5a also shows that the acceptance rate of our approach is around 10%, which is lower than ca. 20% of the
235   formal likelihood function RMSE. Similarly, the convergence rate of our approach is slower than the formal likelihood function (Fig. 5b). It suggests that using the formal likelihood function has a higher efficiency than the approach adapting KGE for calibrations of a system that only contains little uncertainty (only small observation errors in our case). However, when more uncertainties appear, e.g. uncertainties in forcing data and model structures, the convergence rates (efficiency) become similar between the adapted KGE and the formal likelihood function (refer to the following two subsections, Fig. 6b
240   and 8b). Compared to the width of the discharge uncertainty bound using the original KGE (Fig. 5c), calibration using the formal likelihood function (RMSE) reduces the average width of its uncertainty bound by 99.6%, while using the adapted KGE reduces this by 99.5%. Since this virtual experiment does not assume uncertainty in the input data and the model structure, our approach shows a similar performance in the uncertainty estimation as using the formal likelihood function; both can closely reproduce observations.
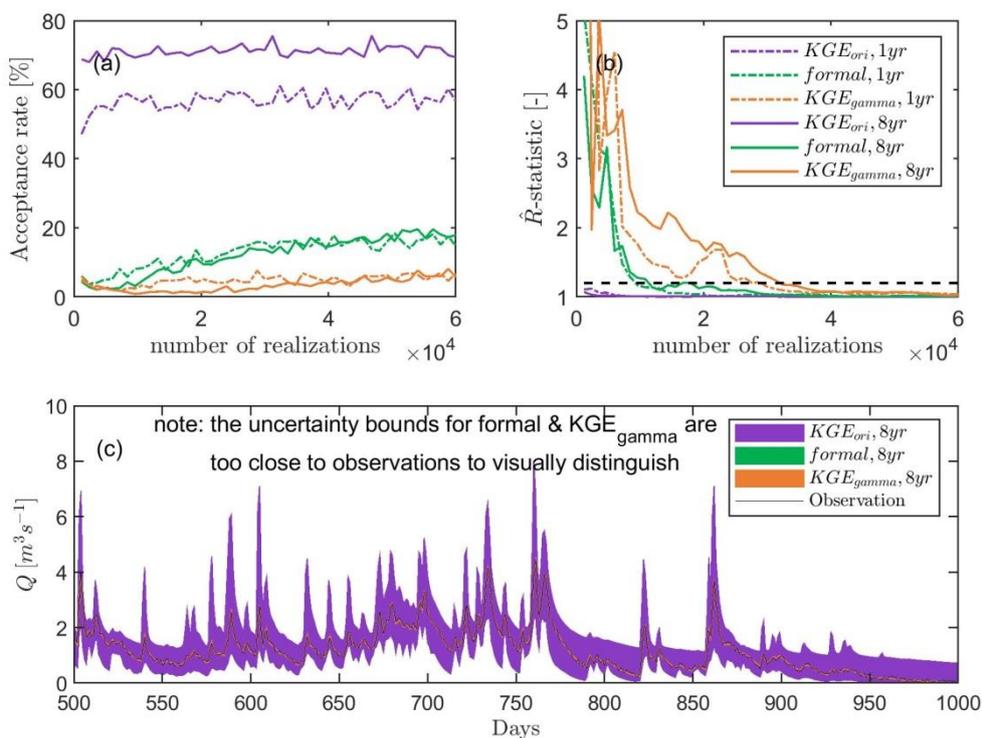
Figure 5 Acceptance rate (a), convergence rate shown with $\widehat{R}$-statistic (b), and uncertainty of discharge simulations (only parameter uncertainty) for the virtual experiment (c). KGE_{ori} indicates using the original KGE as the likelihood function, while KGE_{gamma} represents our new approach using gamma distribution and KGE to derive the probability density, and the case *formal* means using the traditional formal likelihood function (RMSE). The subscripts *1yr* and *8yr* denote calibrations with 1-year and 8-year observations, respectively. Fig. 5c only shows the uncertainty using all data to do calibration to represent the difference between three approaches and shows only 500 days of the simulation to see more detail. The calibrations using less data show similar patterns as Fig. 5c. Note the uncertainty bounds for *formal* and KGE_{gamma} are too small to be visually seen.

3.2 Case study 2: model parameter calibrations with long observations

Calibration for a real system (with uncertainties in forcing, observations, and model structure and parameters) using the original KGE shows a similar trend compared with the calibration to the virtual experiment no matter how much data is included in calibrations (Fig. 6). The acceptance rate of the original KGE is over 80% and it converges very fast such that the uncertainty of simulated discharge is very large (refer to the performance in Fig. 7). Comparing the adapted KGE and the formal likelihood function RMSE, the acceptance rate and convergence rate both decrease with the increase in the amount of observations for calibrations (Fig. 6a and 6b). But to be noticed, the difference of the acceptance rate and convergence rate between the adapted KGE and the formal likelihood function is very small; in particular, the convergence rate is almost identical between the two approaches (Fig. 6b). This indicates that our approach has a same efficiency as the formal likelihood function for calibrations of a system with more uncertainties. From Fig. 6c-6j, we see that using the original KGE

as likelihood function the parameters are difficult to be identified, even with the large increase in the amount of observations

in calibrations. However, with more observations in calibrations, the unidentified parameters *K0* and *UZL* (Fig. 6c and 6d)

265   using the adapted KGE and the formal likelihood function become identified (Fig. 6g and 6h). The identified parameter

values for *K0* (Fig. 6g) show a similar distribution, while the identified parameter values for *UZL* (Fig. 6h) differ between

the adapted KGE and the formal likelihood function. The identified parameter *K1* with 1-year observations in calibration

(Fig. 6e) shows a similar distribution as using 10-year observations (Fig. 6i) between the adapted KGE and the formal

likelihood function. The difference is that the density is higher at the peak when using more observations (Fig. 6i). For the

270   identified parameter *MAXBAS* between our approach and the formal likelihood function, the distribution is similar when

using 1-year observations for calibration (Fig. 6f), but it changes after adding more observations into calibration (Fig. 6g).

This suggests using different likelihood functions may lead to different identified model parameters for a system with

various uncertainties. More information may be needed to confine the model parameters.
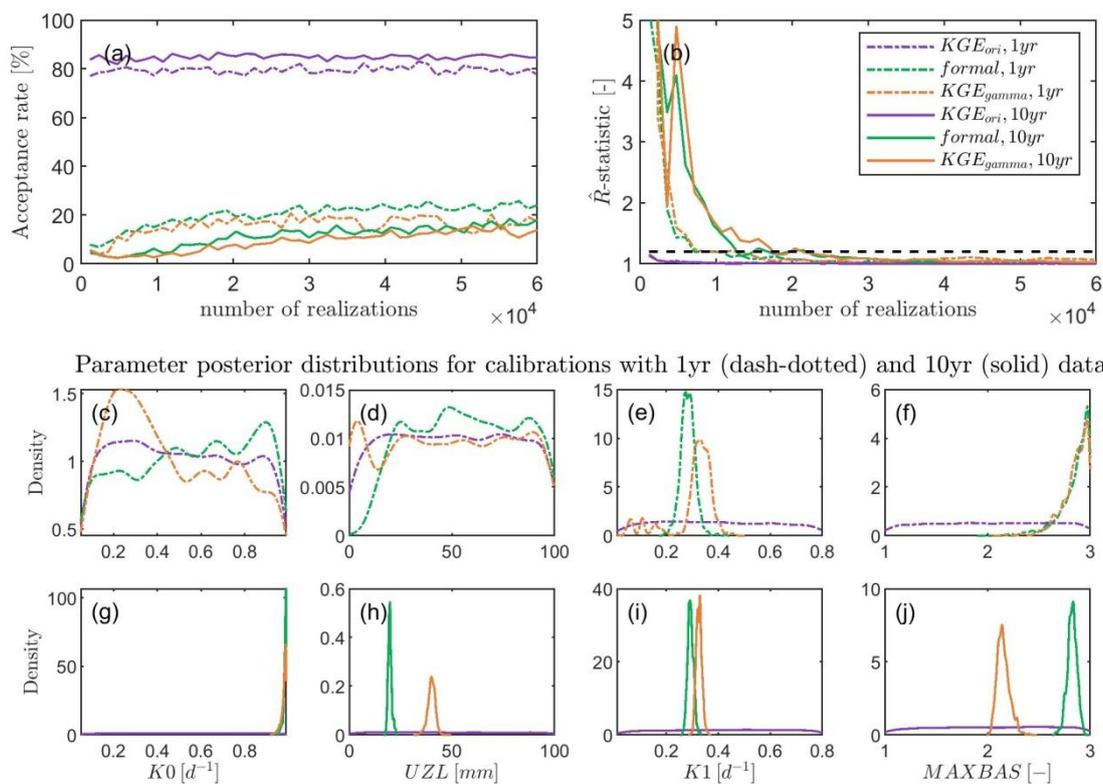

275   **Figure 6** Acceptance rate (a), convergence rate shown with $\hat{R}$-statistic (b), and posterior distributions of selected parameters for calibrations with 1-year (c-f) and 10-year (g-j) observations, respectively. *KGE$_{ori}$* indicates using the original KGE as the likelihood function, while *KGE$_{gamma}$* represents our new approach using gamma distribution and KGE to derive the probability density, and the case *formal* means using the traditional formal likelihood function (RMSE). The subscripts *1yr* and *10yr* denote calibrations with 1-year and 10-year observations, respectively. The four parameters (nine parameters in total) are selected to represent different cases from

280   unidentified (less observations for calibrations) to identified (more observations for calibrations) parameter and show how the identified parameters change when using different amount of observations for calibrations.

In this section, we focus on analyzing the performance in the evaluation period to show the prediction ability of the three approaches. Generally, the uncertainty of the model performance (represented by the interquartile of KGE, α, β and r) of the original KGE is much larger than the adapted KGE or the formal likelihood function (Fig. 7) regardless of total flow, low

285   flow or high flow. With the increasing amount of observations added to calibrations, the performance of the original KGE does not change significantly for all four metrics and for all flow conditions, while using the adapted KGE or the formal likelihood function we can see an increasing trend of the model performance. In the following, we focus on comparing the performance between the adapted KGE and the formal likelihood function. For the total flow, the general performance KGE of our approach is higher than using the formal likelihood function. The two approaches perform similarly for the variability

290   (α) when the amount of calibration data is smaller than 5 years, while our approach tends to overestimate and the formal likelihood function underestimates variability when more data is added to calibrations. Our approach has a smaller overestimation of bias (β) than the formal likelihood function. The two approaches have similar performance in terms of the correlation (r) for the total flow. For the low flow, the performances (all metrics) of all three approaches are poor. Our approach has a better performance in KGE, a lower overestimation of variability and a better simulation of bias, while the

295   formal likelihood function has a better performance in correlation. For the high flow, the adapted KGE and the formal likelihood function perform similarly in terms of KGE, bias (β) and correlation (r). The adapted KGE has a slightly better representation of variability (α) than the formal likelihood function.
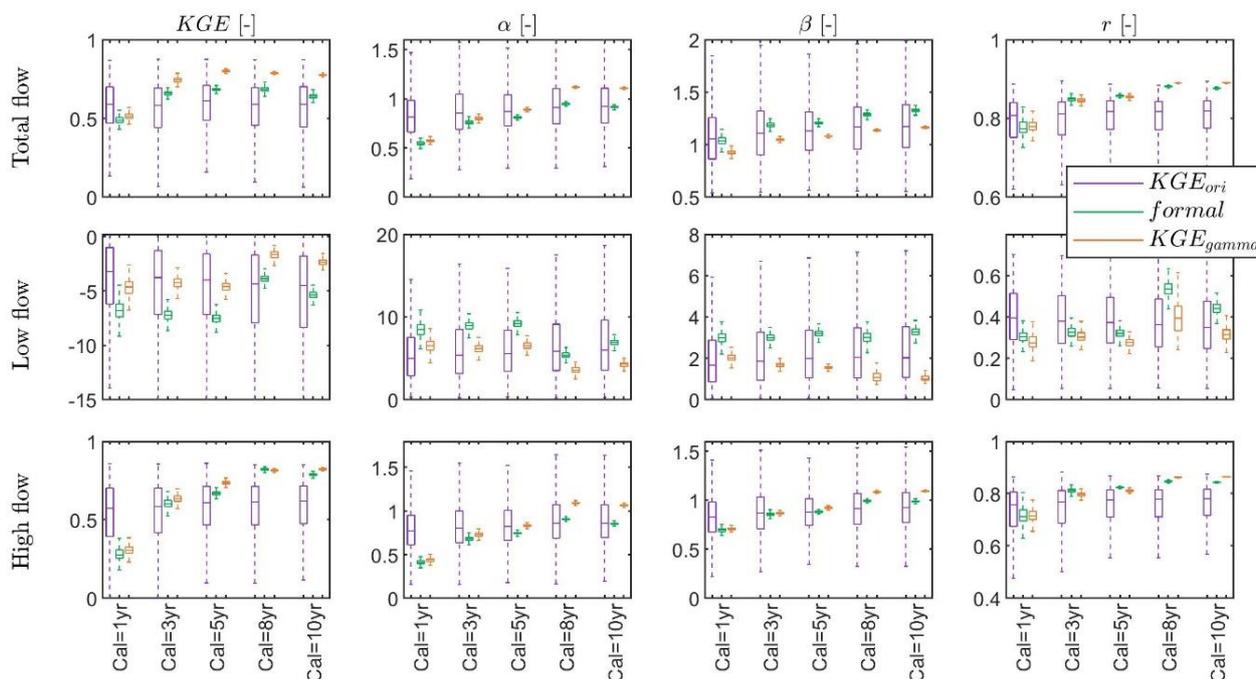
Figure 7 General performance (KGE), variability (α), non-scaled bias (β) and correlation (r) for total flow, low flow (smaller than 10th percentile of observed discharge) and high flow (larger than 90th percentile of observed discharge) during the evaluation period using the original KGE (KGE_{ori}) as the likelihood function, the formal likelihood function RMSE, and our approach using KGE and gamma distribution to derive probability density (KGE_{gamma}) with varying amount of observations (1-year to 10-year) in calibration, for instance, calibration with 1-year observations is shown as Cal=1yr. The boxplot shows the performance of the last 25% of all simulations, which is used to approximate the "true" system behavior in DREAM$_{(ZS)}$. The performance is only shown for the evaluation period to avoid too much information in the boxplot and to represent the prediction ability of different approaches. The performance of the calibration period is provided in Fig. S1 in the supplement. The optimal value for KGE, α, β and r is one, and the closer to one the better the performance.

3.3 Case study 3: model parameter calibrations for a heterogeneous karst system

Fig. 8 shows the acceptance rate, convergence rate and posterior distributions of selected model parameters for a heterogeneous karst system that only has a small amount of observations for model calibration and evaluation. Similar to the case study 1 and 2, the approach using the original KGE has an acceptance rate about 80% and a very fast convergence rate (Fig. 8a and 8b). The acceptance rate for the formal likelihood function is ca. 20% and the adapted KGE about 10% (Fig. 8a). Even though our approach has a smaller acceptance rate than the formal likelihood function for this complex system with short available observations, the convergence rates of both approaches are similar, indicating similar calibration efficiency for a complex system. The posterior distributions of selected model parameters using the original KGE as the likelihood function show a similar behavior as in the case study 1 and 2 that they do not show identified parameter values (Fig. 8c-8f). Between the adapted KGE and the formal likelihood function, we can see a same insensitive parameter a_{fsep} (Fig. 8c) and a same identified parameter $V_{mean,S}$, where the shape of the posterior distributions are similar. In the other hand, we can also

15

Hydrology and
Earth System
Sciences
Discussions
EGU
Open Access

see identified parameters a_{SE} (Fig. 8e) and $V_{mean,E}$ (Fig. 8f) have different posterior distributions. The posterior distribution of the parameters a_{SE} has some overlaps but density peaks are located at different parameter values between our approach and

320   the formal likelihood function. But the posterior distribution of the identified parameter $V_{mean,E}$ shows different shapes between two approaches. This indicates there are interplays for some model processes such that there are compensate of one parameter for another. This should be confined with more prior information.
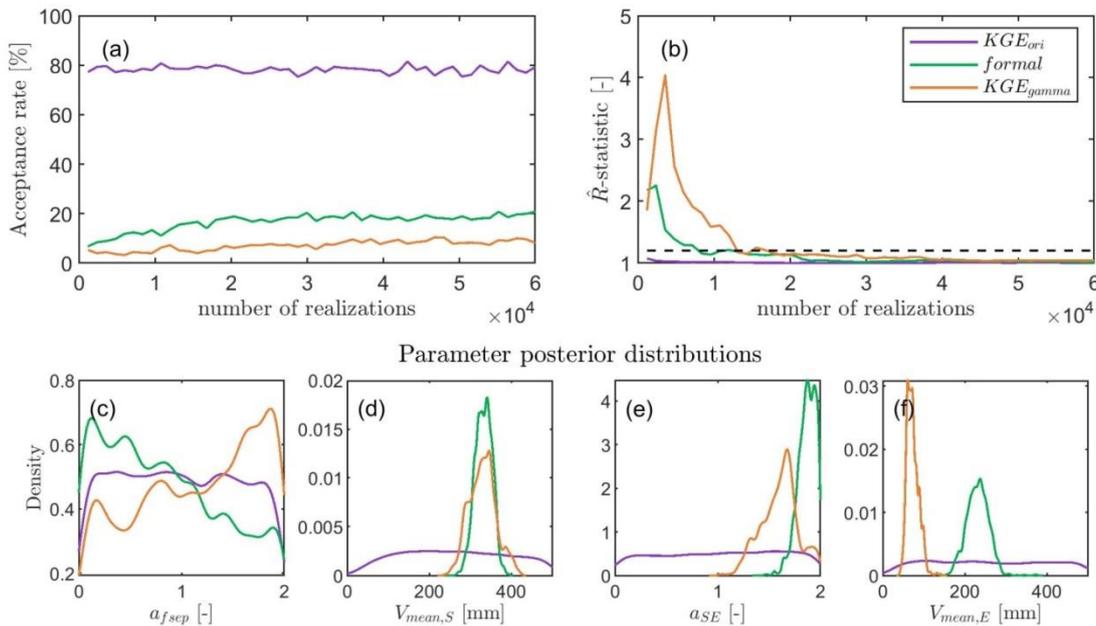

Figure 8 Acceptance rate (a), convergence rate shown with $\hat{R}$-statistic (b), and posterior distributions of selected parameters for
325   calibrations (c-f) for a heterogeneous karst system, respectively. KGE_{ori} indicates using the original KGE as the likelihood function, while KGE_{gamma} represents our new approach using gamma distribution and KGE to derive the probability density, and the case *formal* means using the traditional formal likelihood function (RMSE). The four parameters (seven parameters in total) are selected to represent how different are the posterior distributions of unidentified and identified parameters between the three approaches.

Similar to case study 1 and 2, the uncertainty of the model performance (the interquartile of KGE, α, β and r) of the original
330   KGE is much larger than the other two approaches for all flow conditions. The mean performance represented by the mean values of all four metrics is much lower when using the original KGE as the likelihood function (Fig. 9) for both calibration and evaluation periods at this heterogeneous karst system with short observation records. Therefore, we compare the performance using the adapted KGE and the formal likelihood function. For the total flow, our approach has a higher KGE and a higher performance of variability (α) for both calibration and evaluation periods. The adapted KGE and the formal
335   likelihood function perform similarly in terms of bias (β) and correlation (r). For the low flow and high flow, the adapted KGE and the formal likelihood function both have a better performance in the evaluation period. For the low flow, the general performance KGE of both approaches in the evaluation period is similar, but is low. The overestimation of variability of the adapted KGE is smaller than the formal likelihood function. Our approach has a smaller overestimation of

bias in the calibration period than the formal likelihood function, but in the evaluation period the formal likelihood function

340 has a better performance regarding bias. The correlation in the evaluation period between the adapted KGE and the formal likelihood function is similar. For the high flow, the KGE of the formal likelihood function is slightly higher in calibration, but lower in evaluation compared to our approach. In both calibration and evaluation periods, the two approaches have similar performance in terms of variability and correlation, while our approach has a smaller underestimation of bias compared to the formal likelihood function.
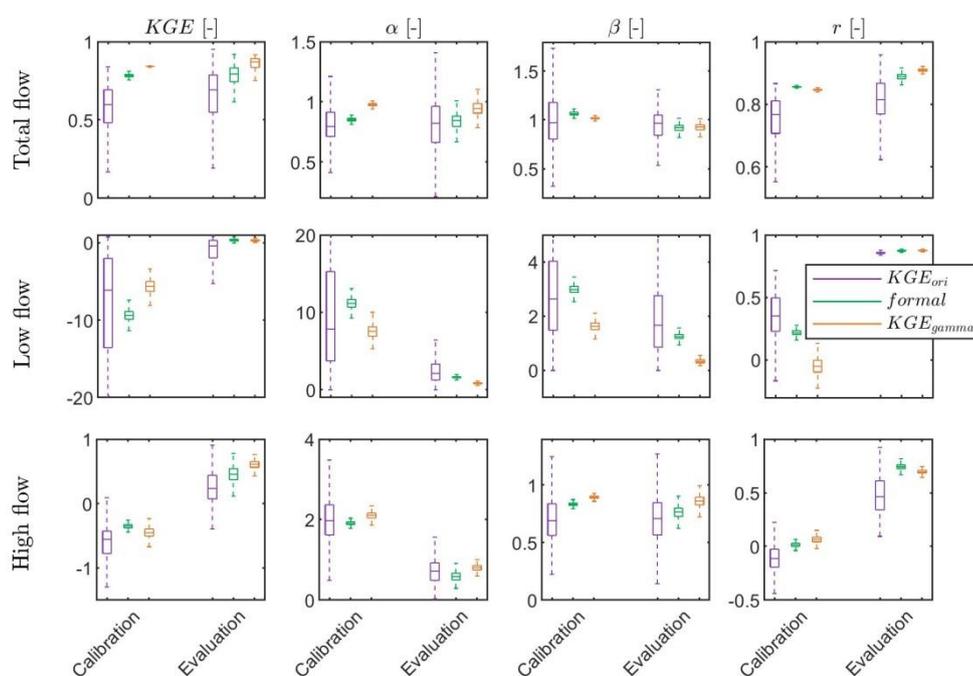
Figure 9 General performance (KGE), variability (α), non-scaled bias (β) and correlation (r) for total flow, low flow (smaller than 10th percentile of observed discharge) and high flow (larger than 90th percentile of observed discharge) for a heterogeneous karst system during calibration and evaluation periods using the original KGE (KGE_{ori}) as the likelihood function, the formal likelihood function RMSE, and our approach using KGE and gamma distribution to derive probability density (KGE_{gamma}). The boxplot shows the performance of the last 350 25% of all simulations, which is used to approximate the "true" system behavior in DREAM$_{(ZS)}$. The optimal value for KGE, α, β and r is one, and the closer to one the better the performance.

4 Discussion

Using the original KGE as the likelihood function for the virtual and real systems, most model parameters are not well identifiable. This results in a very large uncertainty of the discharge simulations. This is because directly using the original

355 KGE as the likelihood estimate assumes a linear increase of probability density with the linear increase of KGE. This leads to the identification of parameter proposals with good KGE performance more difficult and inefficient. The difference

between large and small KGE values is not distinctly large enough that the probability to accept poor proposals is high. This is why we find a very large acceptance rate and a very fast convergence rate. Mantovan and Todini (2006) and Stedinger et al. (2008) also mentioned that using the informal likelihood function, such as Nash-Sutcliffe Efficiency (NSE), in the

360 generalized likelihood uncertainty estimation (GLUE) as objectives cannot find proper posterior distributions of model parameters. Therefore, directly using the original KGE should be avoided and some adaptations to solve the incapability of exploring the posterior distributions such as our approach are needed in MCMC methods.

Our approach can estimate the posterior distributions of model parameters as good as the formal likelihood function in case study 1. This suggests that it is capable to explore the parameter posterior distributions. The adapted KGE has a lower

365 acceptance rate and convergence rate compared to the formal likelihood function for the virtual experiment (case study 1). The possible reason is that one KGE value can cover multiple error combinations with the same RMSE around the true optimum, which makes the RMSE a bit more efficient to draw proposals for parameters very close to the true optimum (known parameters in case study 1) for a system that only contains little uncertainty. However, calibrations of real systems usually contain more uncertainties e.g., uncertainties in forcing (including the spatial averaging), observation data

370 (measurement errors), and uncertainty in model structures. The adapted KGE has a similar convergence rate (efficiency) as the formal likelihood for the real-world calibrations (case study 2 and 3). In particular, the acceptance rate of the adapted KGE is around 20% for a system that we have good input and observations (case study 2). This is similar to the formal likelihood function and is also close to the theoretically optimal acceptance rate (0.234) in Metropolis algorithms with random walk (Yang et al., 2020).

375 The uncertainty bound of discharge simulations in case study 1 is almost identical between the adapted KGE and the formal likelihood function. This indicates our approach can behave similarly concerning discharge uncertainty estimation as the formal likelihood function. For the calibration to the real systems, our approach even has a higher general performance in terms of the mean KGE of the evaluation for the total, low and high flows. McMillan & Clark (2009) had a similar finding that using another informal likelihood function, NSE, in MCMC methods outperforms the formal likelihood for calibrations

380 with high variability and multi-optimum. The formal likelihood functions go along with the strong assumption that errors are distributed normally (Vrugt et al., 2008, 2009), the informal likelihood function KGE takes into account more variability without strict assumptions on error sources (Gupta et al., 2009). The adapted KGE performs similar to the formal likelihood function regarding the correlation between simulations and observations shown in case study 2 and 3. However, the two approaches both have lower performance for the low flow simulations compared to the total and high flow. Our approach

385 shows a lower overestimation of bias in low flows than the formal likelihood function shown as the metric β in case study 2 and 3. Jeannin et al. (2021) found that using formal likelihood functions such as RMSE as objectives has a large bias in baseflow simulations. For RMSE as the objective, each individual error has the same weight. To have a high overall performance, the optimization tends to firstly fit the high flows since its error is relatively larger and the contribution to RMSE is thus larger. The adapted KGE and the formal likelihood function both have a better representation of variability (α)

390 with more observations included in calibrations. This makes sense since more data is involved in calibration more information on variability will be captured in calibrations.

While our approach has a similar performance as the formal likelihood function, we find similar posterior distributions for certain parameters but also inconsistent posterior distributions for some parameters between the two approaches. This suggests that some model processes interplay with other processes such that there are compensates of one parameter for

395 another, i.e. parameter interactions. Additional information, e.g. solutes, can help to constrain model parameters (Hartmann et al., 2017) to represent the complexity of real hydrological systems. Many studies have shown that multi-objective calibrations allow to adequately and properly estimate important characteristics of a system (Vrugt, et al., 2003; Yapo et al., 1998). Using KGE can provide a feasible way to combine various types of observations as a measure of multi-objective performance and avoids issues regarding data units, scales and frequency.

400 Even though our approach adapts the gamma distribution to compute the probability density for KGE, the way we formulate the likelihood function based on KGE is still informal. It means the derivation of the likelihood is not from a strict theoretical probability framework, which is a limitation of our approach. Nevertheless, our approach provides a feasible and pragmatic way and a close solution to the formal likelihood function to avoid the pitfalls of directly using the original KGE in MCMC methods. Future work is needed to find a solution to link probability density and KGE in order to incorporate KGE in a

405 statistical manner as much as possible.

5 Conclusions

Our study demonstrates that using the original KGE in DREAM$_{(ZS)}$ results in a very high acceptance rate and a large uncertainty bound of discharge simulations. This is due to the confusing evolution orientation for negative KGE values and the nonlinear performance of KGE. To solve these two problems, we propose adapting KGE with the gamma distribution to

410 formulate the pseudo log-likelihood function to avoid negative posterior density ratios and to include a proper nonlinearity of performance. With three case studies we demonstrate that the adapted KGE performs as good as the formal likelihood function for the exploration of the posterior distributions of model parameters. Through the calibrations varying the amount of observations included in the calibration, we show that the adapted KGE is robust as the formal likelihood function. Our approach even has a higher general performance, the mean KGE of the evaluation, and a smaller bias overestimation of low

415 flows than the formal likelihood function. The limitation of our approach is the lack of theoretical probability derivation. Besides that formal limitation, our approach keeps the advantages of KGE, e.g. consideration of variability and possibilities to combine multiple types of data, and performs like a pseudo formal likelihood. Thus, it provides a feasible way to use KGE as an informal likelihood function in MCMC methods.

Data and code availability

420 All data used in this study has been published in Liu, et al. (2021). *Water Resources Research*, *57*(1). https://doi.org/10.1029/2020WR028598; Hartmann, et al. (2014). *Water Resources Research*, *50*(8), 6507–6521. https://doi.org/10.1002/2014WR015685 and the dataset publicly available via https://ral.ucar.edu/solutions/products/camels. The Matlab code on using our approach to calculate the likelihood is provided in the supplement.

Author contribution

425 YL conceptualized the study, developed and applied the adapted KGE approach, and visualized the results. YL and JFO wrote the paper, and analyzed the results. MM and AH provided supervision and advice throughout developing the adapted KGE approach and supported the development of this manuscript. All authors contributed to the revision of the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

References

Addor, N., Newman, A. J., Mizukami, N. and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for
440 large-sample studies, Earth Syst. Sci, 21, 5293–5313, doi:10.5194/hess-21-5293-2017, 2017.

Beven, K. J., Smith, P. J. and Freer, J. E.: So just why would a modeller choose to be incoherent?, J. Hydrol., 354(1–4), 15–32, doi:10.1016/j.jhydrol.2008.02.007, 2008.

Freer, J., Beven, K. and Ambroise, B.: 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, doi:10.1029/95WR03723, 1996.

445 Gupta, H. V., Kling, H., Yilmaz, K. K. and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377(1–2), 80–91, doi:10.1016/j.jhydrol.2009.08.003, 2009.

Hartmann, A., Barberá, J. A., Lange, J., Andreo, B. and Weiler, M.: Progress in the hydrologic simulation of time variant recharge areas of karst systems – Exemplified at a karst spring in Southern Spain, Adv. Water Resour., 54, 149–160, 450 doi:10.1016/j.advwatres.2013.01.010, 2013.

Hartmann, A., Mudarra, M., Andreo, B., Marín, A., Wagener, T. and Lange, J.: Modeling spatiotemporal impacts of hydroclimatic extremes on groundwater recharge at a Mediterranean karst aquifer, Water Resour. Res., 50(8), 6507–6521, doi:10.1002/2014WR015685, 2014.

Hartmann, A., Antonio Barberá, J. and Andreo, B.: On the value of water quality data and informative flow states in karst 455 modelling, Hydrol. Earth Syst. Sci., 21(12), 5971–5985, doi:10.5194/hess-21-5971-2017, 2017.

Jeannin, P.-Y., Artigue, G., Butscher, C., Chang, Y., Charlier, J.-B., Duran, L., Gill, L., Hartmann, A., Johannet, A., Jourde, H., Kavousi, A., Liesch, T., Liu, Y., Lüthi, M., Malard, A., Mazzilli, N., Pardo-Igúzquiza, E., Thiéry, D., Reimann, T., Schuler, P., Wöhling, T. and Wunsch, A.: Karst modelling challenge 1: Results of hydrological modelling, J. Hydrol., 600, 126508, doi:10.1016/j.jhydrol.2021.126508, 2021.

460 Knoben, W. J. M., Freer, J. E. and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores, Hydrol. Earth Syst. Sci., 23(10), 4323–4331, doi:10.5194/hess-23-4323-2019, 2019.

Lindström, G., Johansson, B., Persson, M., Gardelin, M. and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, J. Hydrol., 201(1–4), 272–288, doi:10.1016/S0022-1694(97)00041-3, 1997.

Liu, Y., Wagener, T. and Hartmann, A.: Assessing Streamflow Sensitivity to Precipitation Variability in Karst-Influenced 465 Catchments With Unclosed Water Balances, Water Resour. Res., 57(1), doi:10.1029/2020WR028598, 2021.

Mantovan, P. and Todini, E.: Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, J. Hydrol., 330(1–2), 368–381, doi:10.1016/j.jhydrol.2006.04.046, 2006.

McMillan, H. and Clark, M.: Rainfall-runoff model calibration using informal likelihood measures within a Markov chain Monte Carlo sampling scheme, Water Resour. Res., 45(4), 1–12, doi:10.1029/2008WR007288, 2009.

470 Newman, A., Sampson, K., Clark, M. P., Bock, A., Viger, R. J. and Blodgett, D.: A large-sample watershed-scale hydrometeorological dataset for the contiguous USA. Boulder, CO: UCAR/NCAR. https://dx.doi.org/10.5065/D6MW2F4D, 2014.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T. and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the

475   contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance, Hydrol. Earth Syst. Sci., 19(1), 209–223, doi:10.5194/hess-19-209-2015, 2015.

Pool, S., Vis, M. and Seibert, J.: Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency, Hydrol. Sci. J., 63(13–14), 1941–1953, doi:10.1080/02626667.2018.1552002, 2018.

Smith, T. J. and Marshall, L. A.: Bayesian methods in hydrologic modeling: A study of recent advancements in Markov

480   chain Monte Carlo techniques, Water Resour. Res., 44(12), 1–9, doi:10.1029/2007wr006705, 2008.

Stedinger, J. R., Vogel, R. M., Lee, S. U. and Batchelder, R.: Appraisal of the generalized likelihood uncertainty estimation (GLUE) method, Water Resour. Res., 44(12), 1–17, doi:10.1029/2008wr006822, 2008.

Vrugt, J. A.: Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation, Environ. Model. Softw., 75, 273–316, doi:10.1016/j.envsoft.2015.08.013, 2016.

485   Vrugt, J. A., Gupta, H. V., Bouten, W. and Sorooshian, S.: A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, Water Resour. Res., 39(8), doi:10.1029/2002WR001642, 2003a.

Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W. and Sorooshian, S.: Effective and efficient algorithm for multiobjective optimization of hydrologic models, Water Resour. Res., 39(8), 1–19, doi:10.1029/2002WR001746, 2003b.

490   Vrugt, J. A., Ter Braak, C. J. F., Clark, M. P., Hyman, J. M. and Robinson, B. A.: Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, Water Resour. Res., 44(12), 2008.

Vrugt, J. A., Ter Braak, C. J. F., Diks, C. G. H., Robinson, B. A., Hyman, J. M. and Higdon, D.: Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, Int. J. Nonlinear Sci.

495   Numer. Simul., 10(3), 273–290, 2009.

Yang, J., Roberts, G. O. and Rosenthal, J. S.: Optimal scaling of random-walk metropolis algorithms on general target distributions, Stoch. Process. their Appl., 130(10), 6094–6132, doi:10.1016/j.spa.2020.05.004, 2020.

Yapo, P. O., Gupta, H. V. and Sorooshian, S.: Multi-objective global optimization for hydrologic models, J. Hydrol., 204(1–4), 83–97, doi:10.1016/S0022-1694(97)00107-8, 1998.

500