Hydrol. Earth Syst. Sci., 23, 2601–2614, 2019
https://doi.org/10.5194/hess-23-2601-2019

Hydrology and Earth System Sciences
Open Access
EGU

On the choice of calibration metrics for "high-flow" estimation using hydrologic models

Naoki Mizukami[1], Oldrich Rakovec[2,3], Andrew J. Newman[1], Martyn P. Clark[1,a], Andrew W. Wood[1], Hoshin V. Gupta[4], and Rohini Kumar[2]

[1]National Center For Atmospheric Research, Research Application Laboratory, Boulder, CO, USA
[2]UFZ-Helmholtz Centre for Environmental Research, Department of Computational Hydrosystems, Leipzig, Germany
[3]Czech University of Life Sciences, Faculty of Environmental Sciences, Prague, Czech Republic
[4]Department of Hydrology and Atmospheric Sciences, University of Arizona, Tucson, AZ, USA
[a]now at: Coldwater Laboratory, University of Saskatchewan, Canmore, Alberta, Canada

Correspondence: Naoki Mizukami (mizukami@ucar.edu)

Received: 19 July 2018 – Discussion started: 3 August 2018
Revised: 18 April 2019 – Accepted: 6 May 2019 – Published: 17 June 2019

Abstract. Calibration is an essential step for improving the accuracy of simulations generated using hydrologic models. A key modeling decision is selecting the performance metric to be optimized. It has been common to use squared error performance metrics, or normalized variants such as Nash–Sutcliffe efficiency (NSE), based on the idea that their squared-error nature will emphasize the estimates of high flows. However, we conclude that NSE-based model calibrations actually result in *poor* reproduction of high-flow events, such as the annual peak flows that are used for flood frequency estimation. Using three different types of performance metrics, we calibrate two hydrological models at a daily step, the Variable Infiltration Capacity (VIC) model and the mesoscale Hydrologic Model (mHM), and evaluate their ability to simulate high-flow events for 492 basins throughout the contiguous United States. The metrics investigated are (1) NSE, (2) Kling–Gupta efficiency (KGE) and its variants, and (3) annual peak flow bias (APFB), where the latter is an application-specific metric that focuses on annual peak flows. As expected, the APFB metric produces the best annual peak flow estimates; however, performance on other high-flow-related metrics is poor. In contrast, the use of NSE results in annual peak flow estimates that are more than 20 % worse, primarily due to the tendency of NSE to underestimate observed flow variability. On the other hand, the use of KGE results in annual peak flow estimates that are better than from NSE, owing to improved flow time series metrics (mean and variance), with only a slight degradation in performance with respect to other related metrics, particularly when a non-standard weighting of the components of KGE is used. Stochastically generated ensemble simulations based on model residuals show the ability to improve the high-flow metrics, regardless of the deterministic performances. However, we emphasize that improving the fidelity of streamflow dynamics from deterministically calibrated models is still important, as it may improve high-flow metrics (for the right reasons). Overall, this work highlights the need for a deeper understanding of performance metric behavior and design in relation to the desired goals of model calibration.

1 Introduction

Computer-based hydrologic, land-surface, and water balance models are used extensively to generate continuous long-term hydrologic simulations in support of water resource management, planning, and decision making. Such models contain many empirical parameters that cannot be estimated directly from available observations, hence the need for parameter inference by means of the indirect procedure known as calibration (Gupta et al., 2006). In general, all such models require some degree of calibration to maximize their ability to adequately reproduce the observed dynamics of the system response (e.g., streamflow).

A key decision in model calibration is the choice of performance metric (also known as the "objective function") that measures the goodness of fit between the model simulation and system observations. The performance metric can substantially affect the quality of the calibrated model simulations. The most widely used performance metrics are based on comparisons of simulated and observed response time series, including the mean squared error (MSE), Nash–Sutcliffe efficiency (NSE; a normalized version of MSE), and root mean squared error (RMSE; a transformation of MSE). Many previous studies have examined different variants of these metrics (e.g., see Oudin et al., 2006; Kumar et al., 2010; Pushpalatha et al., 2012; Price et al., 2012; Wöhling et al., 2013; Ding et al., 2016; Garcia et al., 2017), including their application to transformations of the system response time series to emphasize performance for specific flow regimes (e.g., use of logarithmic transformation to target low flows) or using combinations of different metrics to obtain balanced performance on different flow regimes.

As an alternative to metrics that measure the distance between response time series, the class of *hydrologic signature* metrics (e.g., Olden and Poff, 2003; Shamir et al., 2005; Gupta et al., 2008; Yilmaz et al., 2008; Westerberg and McMillan, 2015; Westerberg et al., 2016; Addor et al., 2017a) has been gaining popularity for hydrologic model calibration (Yadav et al., 2007; Westerberg et al., 2011; Shafii and Tolson, 2015; Kavetski et al., 2018). A hydrologic signature is a metric that quantifies a targeted property or behavior of a hydrologic time series (e.g., that of a specific portion such as peaks, recessions, water balance, flow variability, or flow correlation structure), in such a way that it is informative regarding a specific hydrologic process of a catchment (Yilmaz et al., 2008).

The use of hydrologic signatures to form metrics for model calibration requires selection of a full set of appropriate signature properties that are relevant to all of the aspects of system behavior that are of interest in a given situation. As discussed by Gupta et al. (2008), the use of multiple hydrologic signatures for model calibration involves the use of multi-objective optimization (Gupta et al., 1998) in which a trade-off among the ability to optimize different signature metrics must be resolved. This means that, in the face of model structural errors, it is typically impossible to simultaneously obtain optimal performance on all of the metrics (in addition to the practical difficulty of determining the position of the high-dimensional Pareto front). Further, if only a small subset of signature metrics is used for calibration, the model performance in terms of the non-included metrics can suffer (Shafii and Tolson, 2015). The result of calibration using a multi-objective approach is a Pareto set of parameters, where different locations in the set emphasize different degrees of fit to the different hydrological signatures.

In general, water resource planners focus on achieving maximum accuracy in terms of specific hydrologic properties and will therefore select metrics that target the requirements of their specific application while accepting (if necessary) reduced model skill in other aspects. For example, in climate change impact assessment studies, reproduction of monthly or seasonal streamflow is typically more important than behaviors at finer temporal resolutions, and so hydrologists typically use monthly rather than daily error metrics (Elsner et al., 2010, 2014). Hereafter this metric is referred to as an "application-specific metric". It is worth noting that the application-specific metric can be a hydrologic signature metric. For example, high-flow volume based on the flow duration curve characterizes the surface flow processes and may be of interest for estimation of flood frequency.

In this study, we examine how the formulation of the performance metric used for model calibration affects the overall functioning of system response behaviors generated by hydrologic models, with a particular focus on high-flow characteristics. The specific research questions addressed in this paper are the following.

1. How do commonly used time-series-based performance metrics perform compared to the use of an application-specific metric?

2. To what degree does use of an application-specific metric result in reduced model skill in terms of other metrics not directly used for model calibration?

We address these questions by studying the high-flow characteristics and flood frequency estimates for a diverse range of 492 catchments across the contiguous United States (CONUS) generated by two models: the mesoscale Hydrologic Model (mHM; Kumar et al., 2013b; Samaniego et al., 2010, 2017) and the Variable Infiltration Capacity (VIC; Liang et al., 1994) model. Our focus on high-flow estimation is motivated by (a) their importance to a wide range of hydrologic applications related to high-flow characteristics (e.g., flood forecasting, flood frequency analysis) and their relevance to historical change and future projections (Wobus et al., 2017); and (b) persistent lack of community-wide awareness of the pitfalls associated with use of squared error type metrics for high-flow estimation. Specifically, we compared and contrasted the model simulation results of the calibration based on metric (1) NSE, (2) Kling–Gupta efficiency (KGE) and its variants, and (3) annual peak flow bias (APFB) – with a focus on understanding and evaluating the appropriateness of different metrics to capture observed high-flow behaviors across a diverse range of US basins. We also discuss the implications of the choice of different calibration metrics based on stochastic ensemble simulations generated based on remaining model residuals.

The remainder of this paper is organized as follows. Section 2 shows how the use of NSE for model calibration is counter-intuitively problematic when focusing on high-flow estimation. This part of the study is motivated by our experience with CONUS-wide annual peak flow estimates and serves to motivate the need for our large-sample study (Gupta

et al., 2014). Section 3 describes the data, models, and calibration strategy adopted. Section 4 then presents the results followed by discussion in Sect. 5. Concluding remarks are provided in Sect. 6.

2 Motivation

One of the earliest developments of a metric used for model development was by Nash and Sutcliffe (1970), who proposed assessing MSE relative to the observation mean: NSE. A key motivation was to quantify how well the updated model outputs performed when compared against a simple benchmark (the observation mean). Since then, such squared error metrics have been predominantly used for model evaluation as well as for model calibration. Furthermore, MSE-based metrics have been thought to be useful in model calibration to reduce simulation errors associated with high-flow values, because these metrics typically magnify the errors in higher flows more than in the lower flows due to the fact that the errors tend to be heteroscedastic. Although Gupta et al. (2009) showed theoretically how and why the use of NSE and other MSE-based metrics for calibration results in the underestimation of peak flow events, our experience indicates that this notion continues to persist almost a decade later (Price et al., 2012; Ding et al., 2016; Seiller et al., 2017; de Boer-Euser et al., 2017). Via an algebraic decomposition of the NSE into "mean error", "variability error", and "correlation" terms, Gupta et al. (2009) demonstrate that use of NSE for calibration will underestimate the response variability by a proportion equal to the achievable correlation between the simulated and observed responses; i.e., the only situation in which variability is not underestimated is the ideal but unachievable one when the correlation is 1.0. They further show that the consequence is a tendency to underestimate high flows while overestimating low flows (see Fig. 3 in Gupta et al., 2009).

Our recent large-sample calibration study (Mizukami et al., 2017) made us strongly aware of the practical implications of this problem associated with the use of NSE for model calibration. Figure 1 illustrates the bias in the model's ability to reproduce high flows when calibrated with NSE. The plot shows distributions of annual peak flow bias at 492 Hydro-Climate Data Network (HCDN) basins across the CONUS for the VIC model using three different parameter sets determined by Mizukami et al. (2017). Note that the collated parameter set is a patchwork quilt of partially calibrated parameter sets, while the other two sets were obtained via calibration with NSE using the observed data at each basin. The results clearly demonstrate the strong tendency to underestimate annual peak flows at the vast majority of the basins (although calibration at individual basins results in less severe underestimation than the other cases). Figure 1b–d clearly show that annual peak bias is strongly related to variability error but not to mean error (i.e., water balance error). Even though the calibrations resulted in statistically unbiased results over the sample of basins, there is a strong tendency to severely underestimate annual peak flow due to the fact that NSE results in poor statistical simulation of variability. Clearly, the use of NSE-like metrics for model calibration is problematic for the estimation of high flows and extremes. However, improving only simulated flow variability may not improve high-flow estimates in time. It likely also requires improvement of the mean state and daily correlation.

In general, it is impossible to improve the simulation of flow variability (to improve high-flow estimates) without simultaneously affecting the mean and correlation properties of the simulation. To provide a way to achieve balanced improvement of simulated mean flow, flow variability, and daily correlation, Gupta et al. (2009) proposed the KGE as a weighted combination of the three components that appear in the theoretical NSE decomposition formula and showed that this formulation improves flow variability estimates. KGE is expressed as

$$\mathrm{KGE} = 1 - \sqrt{[S_r(r-1)]^2 + [S_\alpha(\alpha-1)]^2 + [S_\beta(\beta-1)]^2},$$
$$\alpha = \frac{\sigma_\mathrm{s}}{\sigma_\mathrm{o}}, \beta = \frac{\mu_\mathrm{s}}{\mu_\mathrm{o}}, \quad (1)$$

where S_r, S_α, and S_β are user-specified scaling factors for the correlation (r), variability ratio (α), and mean ratio (β) terms; σ_s and σ_o are the standard deviation values for the simulated and observed responses, respectively, and μ_s and μ_o are the corresponding mean values. In a balanced formulation, S_r, S_α, and S_β are all set to 1.0. By changing the relative sizes of the S_r, S_α, or S_β weights, the calibration can be altered to more strongly emphasize the reproduction of flow timing, statistical variability, or long-term water balance.

The results of the Mizukami et al. (2017) large-sample study motivated us to carry out further experiments to investigate how the choice of performance metric affects the estimation of peak and high flow. Here, we examine the extent to which altering the scale factors in KGE can result in improved high-flow simulations compared to NSE. We also examine the results provided by use of an application-specific metric, here taken as the percent bias in annual peak flows.

3 Models, datasets, and methods

We use two hydrologic models: VIC (version 4.1.2h) and mHM (version 5.8). The VIC model, which includes explicit soil–vegetation–snow processes, has been used for a wide range of hydrologic applications, and has recently been evaluated in a large-sample predictability benchmark study (Newman et al., 2017). The mHM has been shown to provide robust hydrologic simulations over both Europe and the US (Kumar et al., 2013a; Rakovec et al., 2016b) and is currently being used in application studies (e.g., Thober et al.,

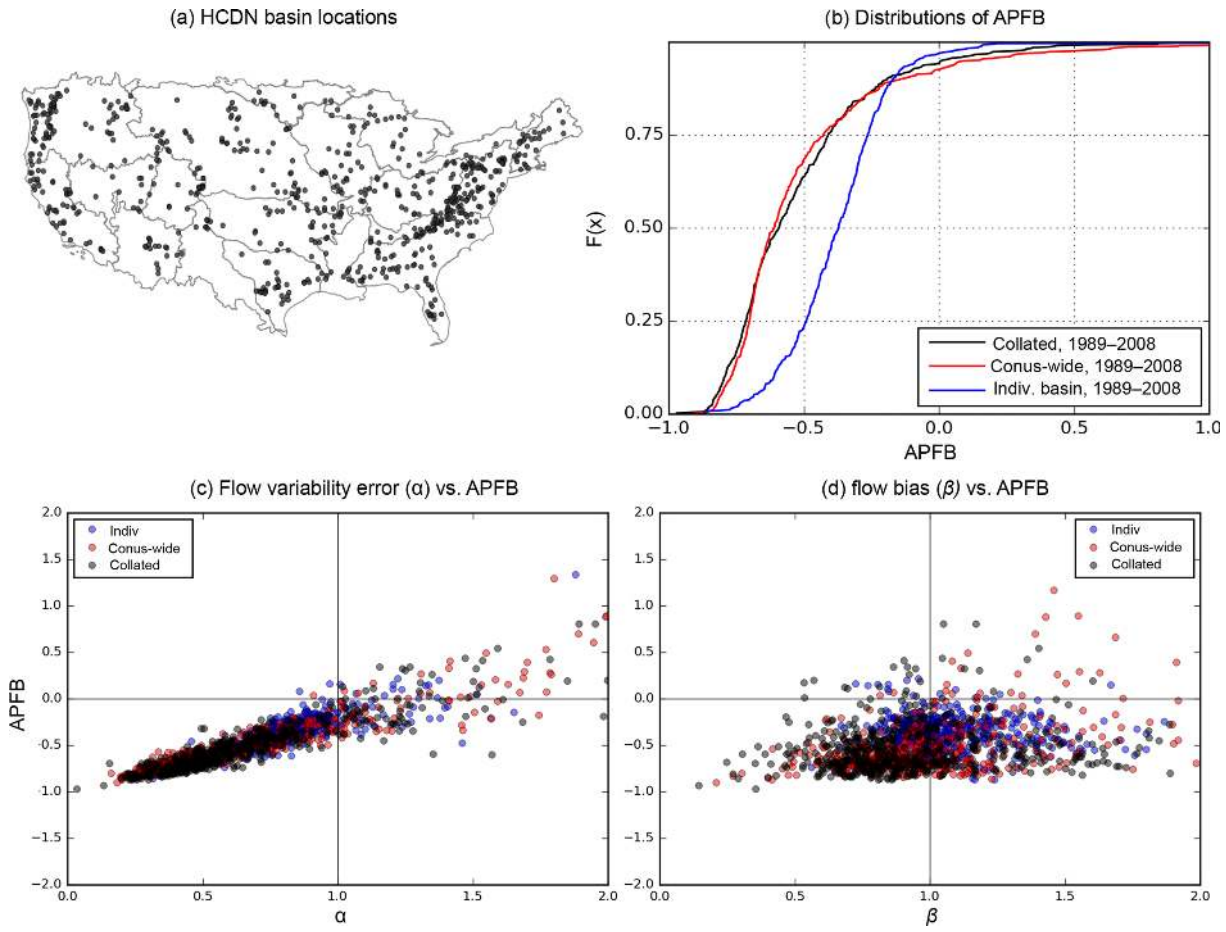

Figure 1. Spatial distribution of Hydro-Climate Data Network (HCDN) basins; **(b)** cumulative distribution of percent bias of annual peak flow (APFB) over 1989–2008 simulated with three different sets of VIC parameters used in Mizukami et al. (2017) at HCDN basins. **(c)** Relationships between variability error (α: simulation-to-observation ratio of daily flow variability) and APFB. **(d)** Relationships between mean error (β: simulation-to-observation ratio of mean flow) and APFB.

2018; Samaniego et al., 2018). We use observed streamflow data at the HCDN basins and daily basin meteorological data from Maurer et al. (2002) for the period from 1980 through 2008, as compiled by the CONUS large-sample basin dataset over a wide range of climate regimes (Newman et al., 2014; Addor et al., 2017b). The use of the large-sample dataset is recommended to obtain general and statistically robust conclusions (Gupta et al., 2014). In the context of flood mechanisms across the CONUS, large flood events are due to precipitation excess in conjunction with antecedent soil moisture states at the majority of the catchments, except that rapid snowmelt events are primarily responsible for floods over the mountainous west (Berghuijs et al., 2016). Both models are run at a daily time step, and each model is calibrated separately for each of the 492 study basins (see Fig. 1a for the basin locations) using several different performance metrics. Although sub-daily simulation is preferable for some flood events, such as flash floods, the effects of the performance metrics on the calibrated high-flow estimates are independent of the simulation time step. Furthermore, instantaneous peak flow (at sub-daily scale) is strongly correlated with daily mean flows (Dieter and Arns, 2003; Ding et al., 2016), justifying daily simulations still providing useful information for instantaneous peak flow estimates. We use a split-sample approach (Klemes, 1986) for the model evaluation. The hydrometeorological data are split into a calibration period (1 October 1999–30 September 2008) and an evaluation period (1 October 1989–30 September 1999), with a prior 10-year warm-up when computing the statistics for each period.

The model parameters calibrated for each model are the same as previously discussed: VIC (Newman et al., 2017; Mizukami et al., 2017) and mHM (Rakovec et al., 2016a, b). Although alternative calibration parameter sets have also been used by others, particularly for VIC (Newman et al., 2017), the purpose of this study is purely to examine the effects of performance metrics used for calibration, and not to obtain "optimal" parameter sets. Each model is identically

configured for each of the 492 basins. Both models use the same set of underlying physiographical and meteorological datasets, so that performance differences can be attributed mainly to the strategy used to obtain the parameter estimates.

Optimization is performed using the dynamically dimensioned search (DDS, Tolson and Shoemaker, 2007) algorithm. Five performance metrics are used for the calibration/evaluation purpose: (1) KGE, (2) KGE-2α, (3) KGE-5α, (4) APFB, and (5) NSE. The first three metrics are KGEs with different scaling factor combinations (S_r, S_α and S_β) = (1, 1, 1), (1, 2, 1), and (1, 5, 1) in Eq. (1), respectively; because variability is strongly correlated with annual peak-flow error (see Fig. 1c), we explore the impact of rescaling the variability error term in Eq. (1). The fourth metric, APFB, is our application-specific high-flow metric, defined as

$$\mathrm{APFB} = \sqrt{[(\mu_{\mathrm{peak}Q_\mathrm{s}}/\mu_{\mathrm{peak}Q_\mathrm{o}} - 1)]^2}, \quad (2)$$

where $\mu_{\mathrm{peak}Q_\mathrm{s}}$ is the mean of the simulated annual peak flow series and $\mu_{\mathrm{peak}Q_\mathrm{o}}$ is the mean of the observed annual peak flow series. Finally, we took NSE as a benchmark performance metric, and compared and contrasted the simulations based on other performance metrics.

The most common choice of KGE scaling factor for hydrologic model calibration has been to set all of them to unity. We applied the KGE in different variants (i.e., with non-unity scaling factors), which to the best of our knowledge have not been studied so far. Note that this scaling is only used to define the performance metric used in model calibration; all performance evaluation results shown in this paper use KGE computed with S_r, S_α, and S_β all set to 1.0.

4 Results

4.1 Overall simulation performance

First, we focus on the general overall performance for the daily streamflow simulations as measured by the performance metrics used. Figures 2 and 3 show the cumulative distributions of the model skill during the evaluation period across the 492 catchments in terms of KGE and its three components: (a) α (standard deviation ratio), (b) β (mean ratio), and (c) r (linear correlation) for VIC (Fig. 2) and mHM (Fig. 3). Considering first the result obtained using KGE, both models, at the median values of the distributions, show improvement in the variability error by approximately 20 % over that obtained using the NSE-based calibration (Figs. 2a and 3a). The plots, however, indicate a continued statistical tendency to underestimate observed flow variability even when the (1, 5, 1) component weighting is used in the scaled KGE-based metric. The corresponding median α and r values obtained for KGE are (α, r) = (0.83, 0.74) for VIC and (α, r) = (0.94, 0.82) for the mHM. Interestingly, the VIC results are more sensitive than the mHM to variations in the S_α weighting. For VIC, the variability estimate continues to improve with increasing S_α (median moves closer to 1.0), but simultaneously leads to overestimation of the mean values (β) and deterioration of correlation (r).

The use of APFB as a calibration metric yields poorer performance for both models, on all of the individual KGE components (wider distributions for α and β, and distribution of r shifted to the left) and consequently on the overall KGE value as well (distribution shifted to the left). In terms of performance as measured by NSE, the use of KGE with the original scaling factors ($\alpha = 1$) results in 3 %–10 % lower NSE than those obtained with the NSE-based calibration case (plots not shown). This is consistent with the expectation that an improvement in the variability error (α closer to unity) leads to deterioration in the NSE score. In general, all the calibration results from both models are consistent with the NSE-based calibration characteristics described in Gupta et al. (2009).

4.2 High-flow simulation performance

Next, we focus on the specific performance of the models in terms of simulation of high flows. As expected, use of the application-specific APFB metric (Eq. 2) leads to the best estimation of annual peak flows for both models (Fig. 4a and b), while use of NSE produces the worst peak flow estimates. Simply switching from NSE to KGE improves APFB by approximately 5 % for VIC and 10 % for the mHM at the median value during the evaluation period. Improvement of APFB occurs at over 85 % of 492 basins for both models. Note that the inter-quartile range of the bias across the basins becomes larger for the evaluation period compared to the calibration period. This is even more pronounced when APFB is used as the objective function (see the results from the mHM; Fig. 4a and b), indicating that the application-specific objective function results in overfitting, and consequently the model is less transferable in time than when the other metrics are used for calibration.

Figure 4c and d show the high-flow simulation performance in terms of another high-flow-related metric – the percent bias in the runoff volume above the 80th percentile of the daily flow duration curve (FHV; Yilmaz et al., 2008). Interestingly, FHV is not reproduced better by the APFB calibrations compared to the other objective functions, particularly for VIC. The implication is that, in this case, the application-specific metric only provides better results for the targeted flow characteristic (here the annual peak flow), but can result in poorer performance for other flow properties (even the closely related annual peak flow). While the mHM model calibrated with APFB does produce a nearly unbiased FHV estimate across the CONUS basins, the inter-quartile range is much larger than that obtained using the other calibration metrics. The VIC model-based results also exhibit larger variability in the FHV bias across the study basins.

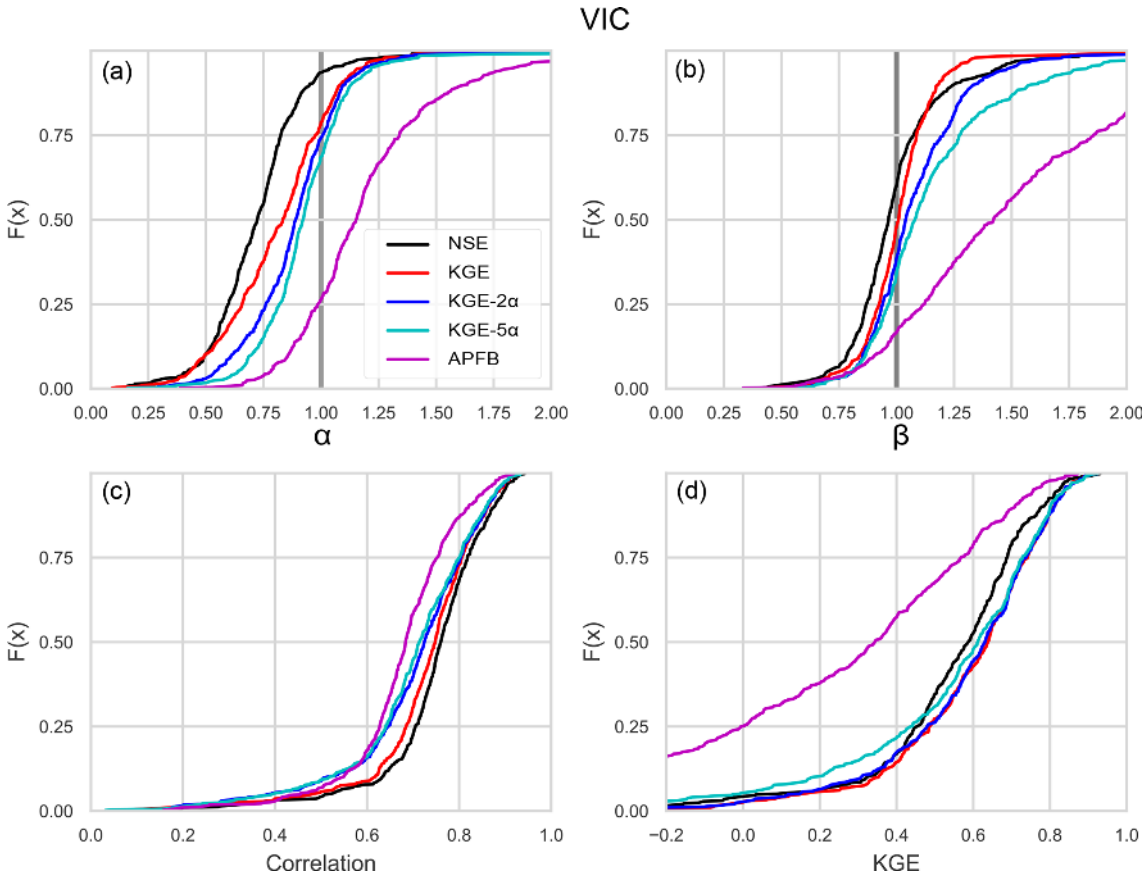

Figure 2. Cumulative distributions of **(a)** flow variability errors α, **(b)** bias β, **(c)** linear correlation r, and **(d)** Kling–Gupta efficiency over the 492 HCDN basin calibrations with five performance metrics for evaluation period and VIC.

4.3 Implication for flood frequency estimation

Annual peak flow estimates are generally used directly in the flood frequency analysis. Figure 5 shows estimated daily flood magnitudes at three return periods (5-,10-, 20-year) using the five different sets of calibration results. Although many practical applications (e.g., floodplain mapping and water infrastructure designs) require estimates of higher extreme events, we focus on a 20-year event (0.95 exceedance probability) for the highest extremes, given use of only 20 years of data for this study; this is to avoid the need for extrapolation of extreme events via theoretical distribution fitting. For this evaluation case (of annual flood magnitudes), we use the combined calibration and evaluation periods.

Figure 5 shows results that are consistent with Fig. 4, although more outlier basins are found to exist for estimates of flood magnitude at the three return periods. The KGE-based calibration improves flood magnitude estimates (compared to NSE) at all three return periods for both models. In particular, mHM especially exhibits a clear reduction of the bias by 10 % at the median compared to the NSE calibration case. The APFB calibration further reduces the bias by 20 % and 10 % for VIC and mHM, respectively. However, regardless of the calibration metric, for both models the peak flows at all return periods are underestimated, although mHM underestimates the flood magnitudes to a lesser degree due to its smaller underestimation of annual peak flow estimates. Even though APFB is less than 5 % at the median value for mHM calibrated with APFB (Fig. 4), the 20-year flood magnitude is underestimated by almost 20 % at the median (Fig. 5). Also, the degree of underestimation of flood magnitude becomes larger with longer return periods.

5 Discussion

While both models show fairly similar trends in skill for each performance metric, it is clear from our large-sample study of 492 basins that the absolute performance of VIC is inferior to that of mHM, irrespective of choice of evaluation metric. A full investigation of why VIC does not perform at the same level of mHM is clearly of interest but is left for future work. To improve the performance of VIC it may be necessary to perform rigorous sensitivity tests similar to comprehensive sensitivity studies that have included investigation of hard-coded parameters in other more complex models (e.g., Men-

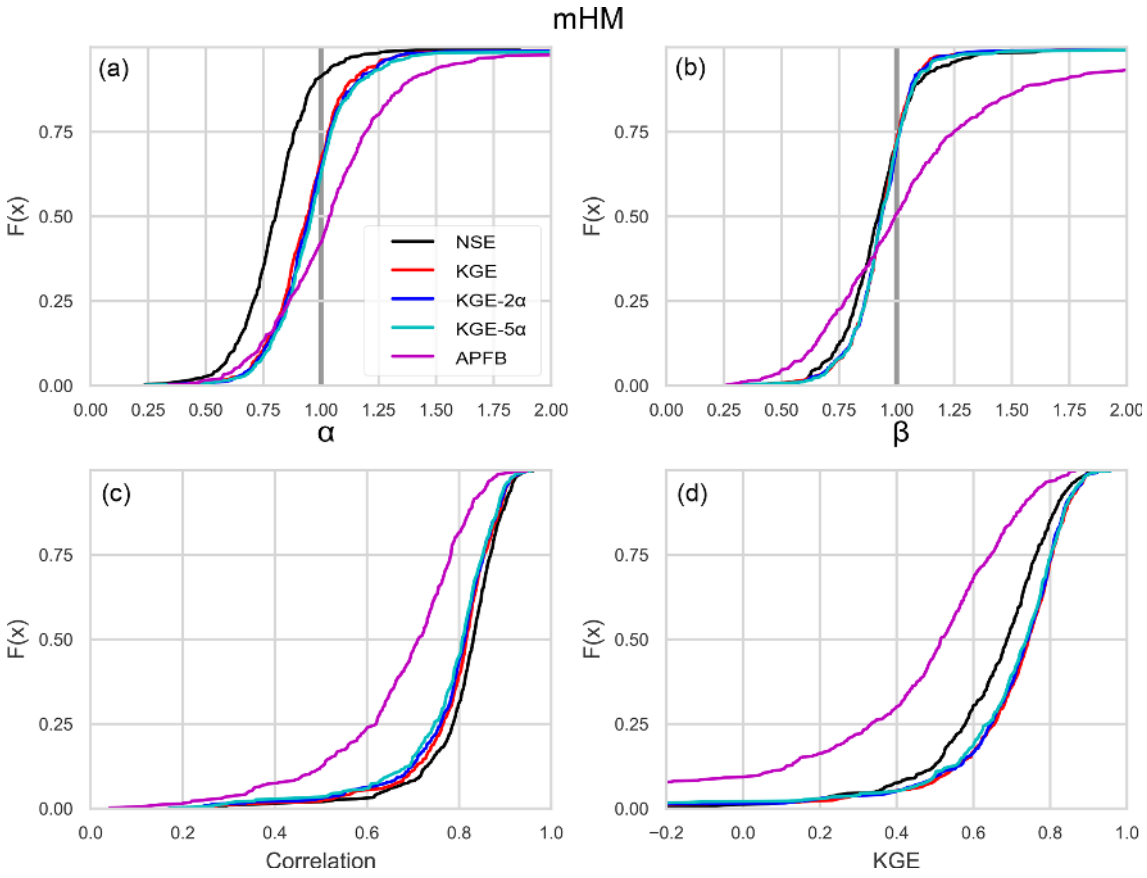

Figure 3. The same as Fig. 2 except for the mHM.

doza et al., 2015; Cuntz et al., 2016). Below, we discuss our results in the context of usage of different performance metrics, in regard to remaining aspects of model errors, and provide suggestions for potential improvement of the high-flow related metrics.

5.1 Consideration of an application-specific metric

Although the annual peak flow estimates improve by switching calibration metrics from NSE to KGE and KGE to APFB, the flood magnitudes are underestimated at all of the return periods examined no matter which performance metric is used for calibration. While the APFB calibration improves, on average, the error of annual peak flow over the 20-year period, the flood magnitude estimates at several percentile or exceedance probability levels are based on estimated peak flow series. Therefore, improving only the bias does not guarantee accuracy of the flood magnitudes at a given return period. Following Gupta et al. (2009), events that are more extreme may be affected more severely by variability errors when examining the series of annual peak flows, particularly because this performance metric accounts only for annual peak flow bias. Figure 6 shows how the estimates of flood magnitudes at the 20-year return period (top panels) and 5-year return period (bottom panels) are related to variability error and bias of annual peak flow estimates. As expected, the more extreme (20-year return period) flood estimates are more strongly correlated with estimates of the variability of annual peak flows than with the 20-year bias of the annual peak flow series. For the less extreme (5-year return period) events, this trend is flipped, and flood magnitude errors are more correlated with the bias.

5.2 Consideration of model residuals

The calibrated models do improve the flow metrics including both time series metrics (mean, variability, etc.) and/or application-specific metrics, depending on the performance metrics used for the calibration. However, residuals always remain after the model calibration because the model never reproduces the observations perfectly. Recently, Farmer and Vogel (2016) discussed the effects of neglecting residuals on estimates of flow metrics, particularly errors in statistical moments of flow time series (mean, variance, skewness, etc.). In the context of this study for the high-flow simulations, let us focus on the flow variability (i.e., variance) component for observation and model simulations, which can expressed by

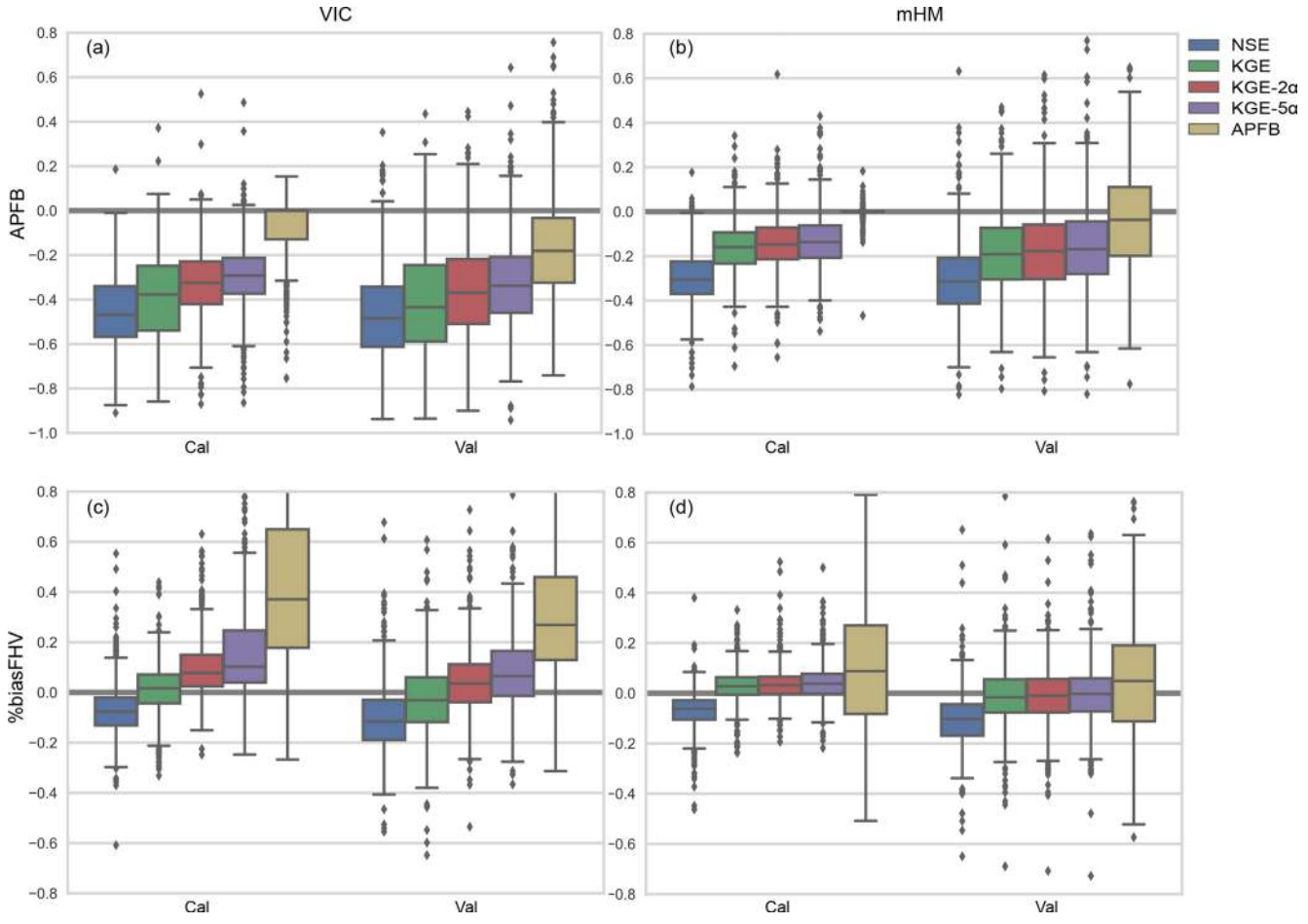

Figure 4. Boxplots of percent bias of APFB (**a**, **b**) and flow volume above the 80th percentile flow duration curve (%biasFHV: **c**, **d**) over the 492 HCDN basin calibrations with five performance metrics for calibration and evaluation periods and two models. Box width represents the inter-quartile range (first and third quartiles), and lower and upper whiskers are placed at 1.5 times the inter-quartile range.

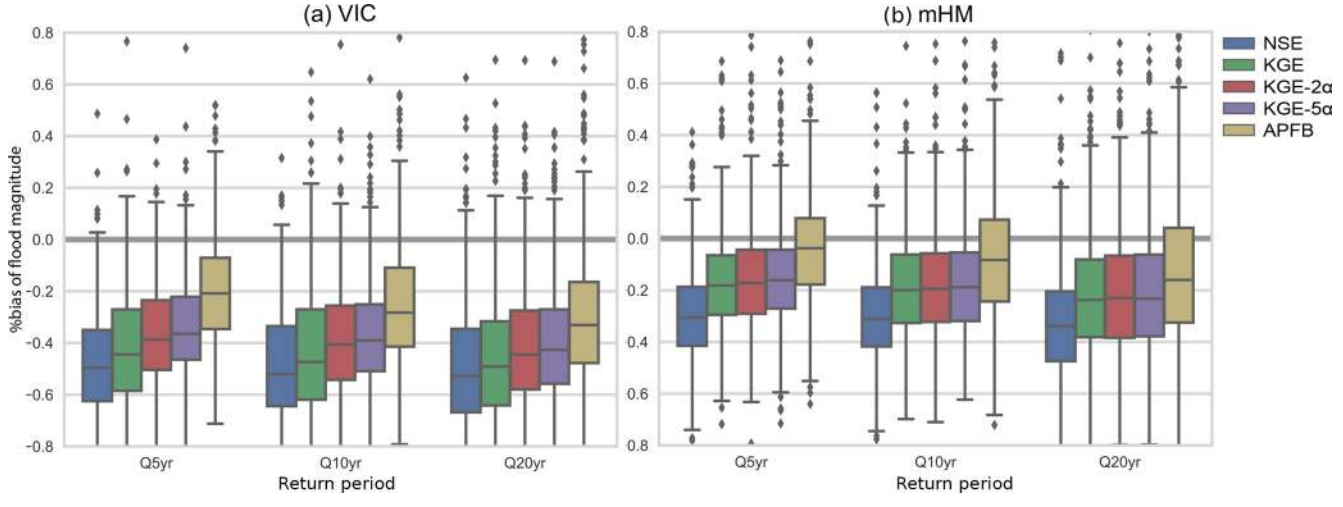

Figure 5. Boxplots of percent bias of flood estimates corresponding to three return periods (5-,10-, and 20-year) over the 492 HCDN basins for the two models. Box-plot representation is the same as Fig. 4.

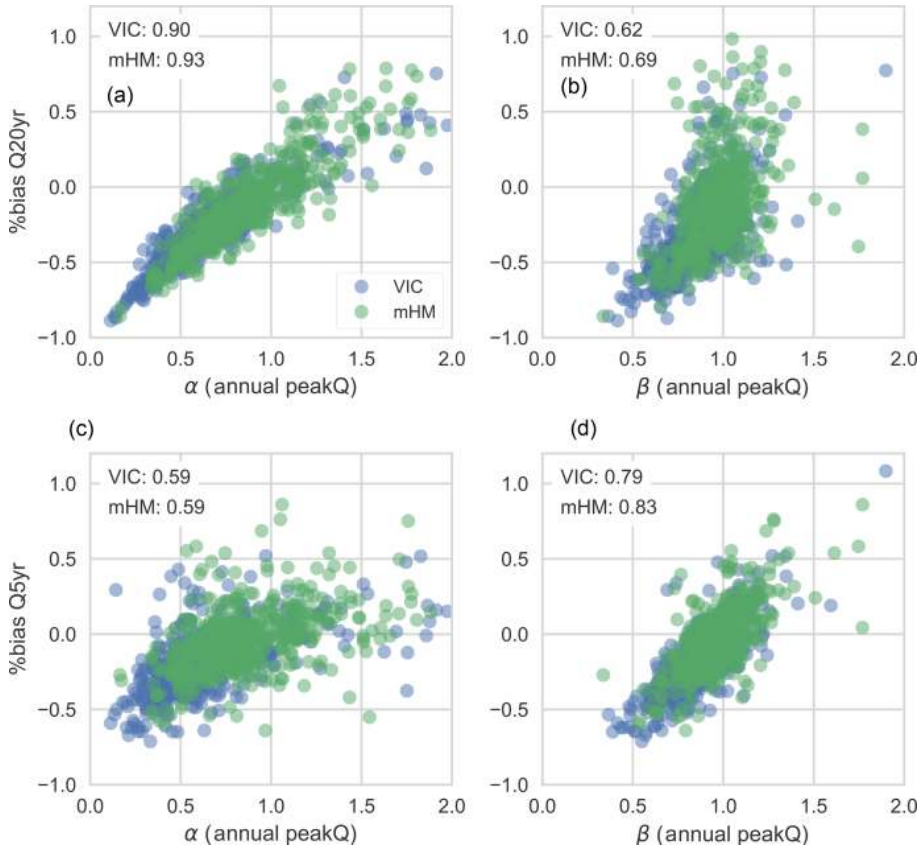

Figure 6. Scatterplots between **(a)** the simulation–observation ratio of variability of annual peak flow series (α) and percent bias of 20-year flood magnitude and **(b)** the simulation–observation ratio of mean annual peak flow series (β) and percent bias of 20-year flood magnitude; **(c)** and **(d)** are the same as **(a)** and **(b)** except for 5-year flood magnitudes. Linear correlations between two variables are specified in the upper-left corner of each plot.

the following equation:

$$\mathrm{Var(o)} = \mathrm{Var(s+\epsilon)} = \mathrm{Var(s)} + \mathrm{Var}(\epsilon) + 2\mathrm{COV(s,\epsilon)}, \quad (3)$$

where $\mathrm{Var}(X)$ is variance of X, $\mathrm{COV}(X, Y)$ is covariance between X and Y, o is the observed time series, s is simulated time series from the calibrated model, and ϵ is the residuals. The observation time series can be expressed as the sum of the model simulation and residual terms (denoted as $\hat{\mathrm{s}} = \mathrm{s} + \epsilon$). As seen in Eq. (3), neglecting the residuals can match the observed variability only when the variance of the residuals is offset by covariance between the simulation and residuals, i.e., $\mathrm{COV(s,\epsilon)}$. Of course, this condition is not fulfilled (in real-world simulation studies). In our calibration results (as discussed above), the observed flow variability is underestimated for both models in the majority of the study basins for nearly all performance metrics used for the calibration (Figs. 2a and 3a).

To gain more insight into this topic, we examine how stochastically generated residuals, once re-introduced to the simulated flows, can affect the performance metrics. We consider three performance metrics for this analysis: NSE, KGE, and APFB. Figure 7 shows the distributions of flow residuals produced by the calibrated models. The APFB calibration that produces the worst temporal pattern of flow time series (the lowest correlation shown in Figs. 2d and 3d) produces wider residual distributions. Following the method of Bourgin et al. (2015) and Farmer and Vogel (2016), 100 sets of synthetic residual time series (ϵ) are stochastically generated by sampling the residuals of the calibrated flow (i.e., simulation during the calibration period) for each model and added to the respective modeled flow during the evaluation period. The method randomly samples the residuals from the residual pool based on the flow magnitude. For each of the 100 residual amended flow series, mean error (β) and variability error (α) are computed, and then median error values are compared with the original deterministic flow error metrics. Figure 8 shows the improvement of bias (α) and variability error (β) regardless of the performance metric or residual distribution characteristics. Similarly to Farmer and Vogel (2016), high-flow volume error (percent bias of FHV) and APFB computed with residual incorporated flow series also improve compared to the deterministic flow series from the calibrated models (Fig. 9).

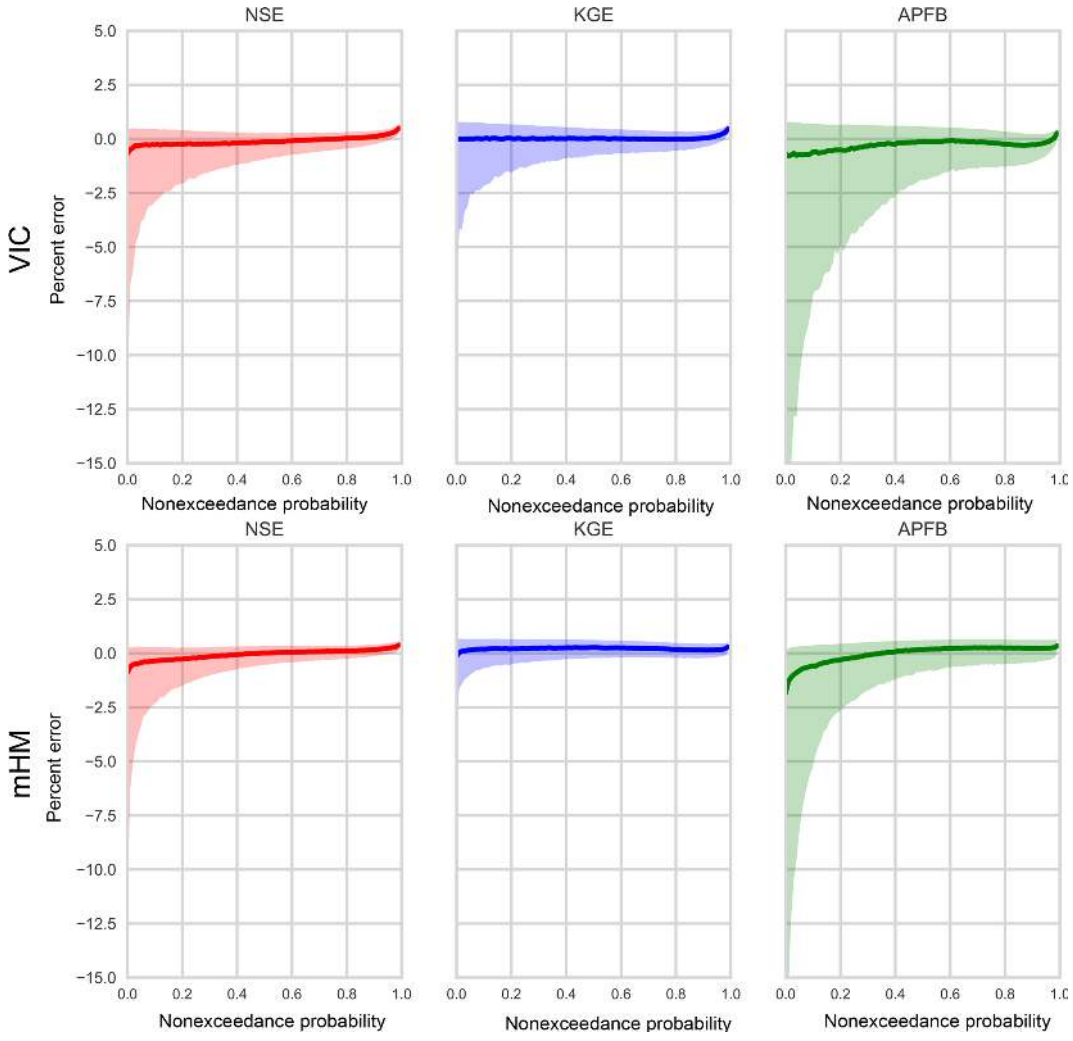

Figure 7. Residual distributions conditioned on the non-exceedance probability of the daily flows over the 492 study basins. Analyses are presented for the three calibration performance metrics. Daily residuals are computed based on the observed and simulated flows during the evaluation period.

The quality of the original deterministic flow simulated by the hydrologic models has little effect on the performance metrics based on the ensemble of residual augmented flows. Since the stochastically generated ensembles do not account for temporal correlation, every ensemble has reduced correlation and deteriorated NSE and KGE metrics. However, the error metric related to the flow duration curve (APFB) is not affected by the lack of correlation because metrics based on the flow duration curve (FDC) do not preserve information regarding the temporal sequence. Although residual augmented flow time series enhances some of the flow metrics, the (temporal) dynamical pattern is not reproduced. These observations point toward the need for careful investigation in interpreting the improvement in model skill, especially when different error metrics are considered.

A key issue is the extent to which high flows are represented in the deterministic and stochastic components. While it is possible to generate ensembles through stochastic simulation of the model residuals (as is done here), and these stochastic simulations improve high-flow error metrics, we will naturally have more confidence in the model simulations if the high flows are well represented in the deterministic model simulations. The use of squared error metrics simply means that a larger part of the high-flow signal must be reconstructed via stochastic simulation.

6 Conclusions

The use of large-sample catchment calibrations of two different hydrologic models with several performance metrics enables us to make robust inferences regarding the effects of the calibration metric on the ability to infer high-flow events. Here, we have focused on improving the representation of

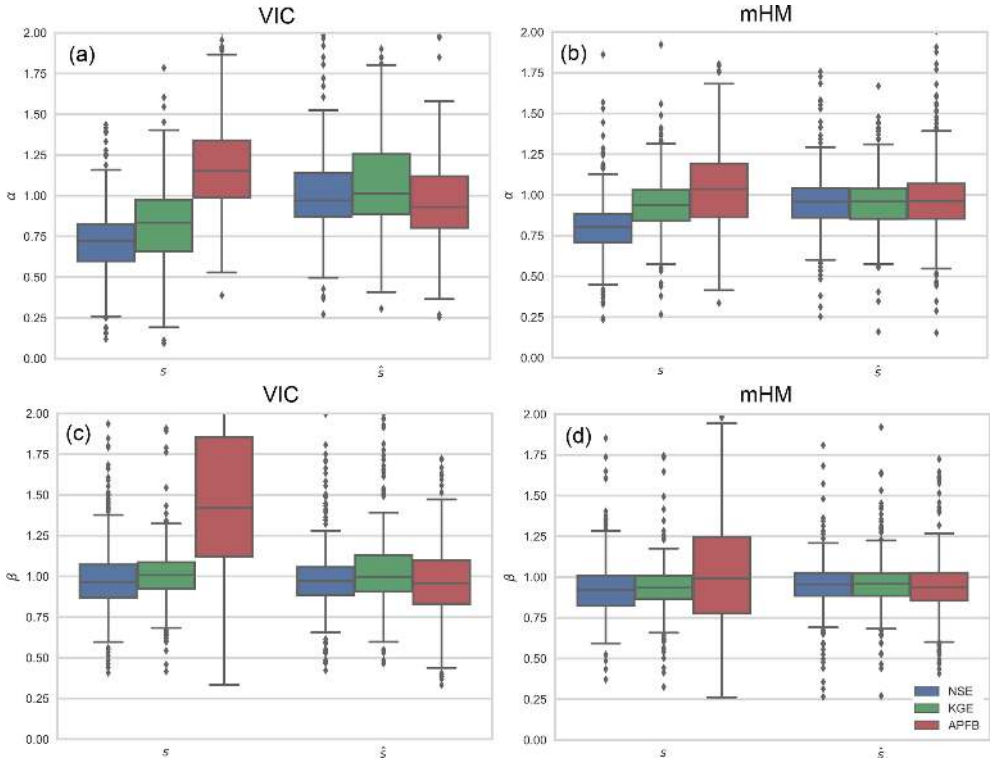

Figure 8. Distribution of the two error metrics (**a**, **b**: α and **c**, **d**: β) computed based on the simulations from NSE-, KGE-, and APFB-calibrated models (labeled as s). The distribution of median error metrics (labeled as ŝ) are based on 100 residual augmented flow series. The evaluation results shown here correspond to the evaluation period. Box-plot representation is the same as Fig. 4.

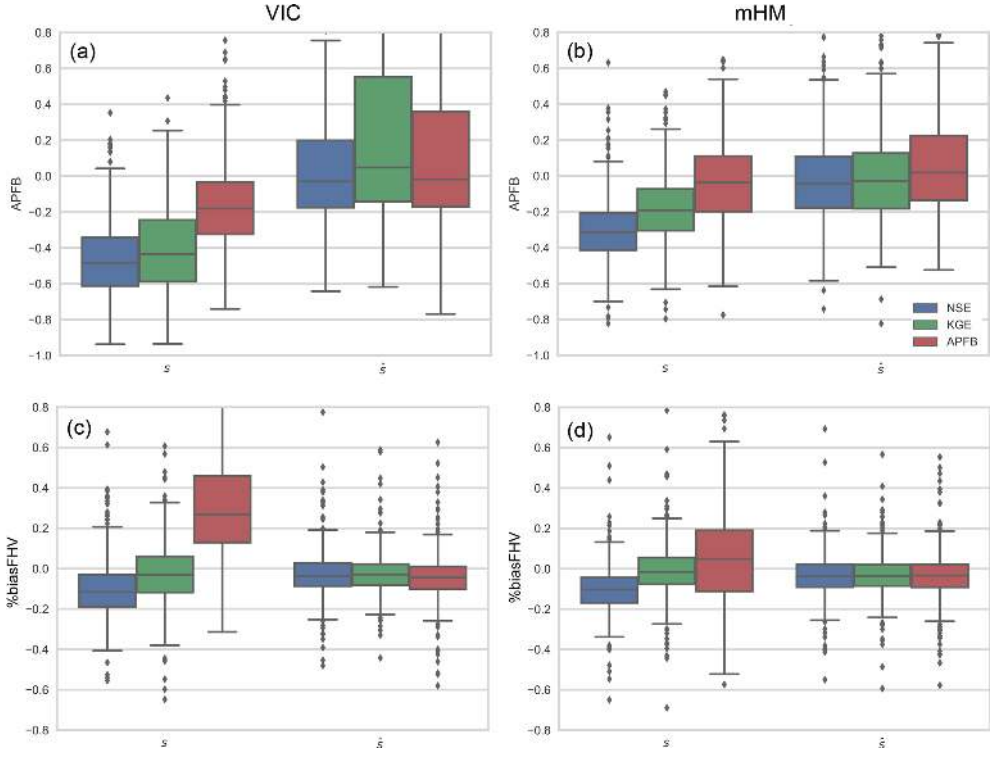

Figure 9. The same as Fig. 8 except for APFB **(a, b)** and percent bias in FHV **(c, d)**.

annual peak flow estimates, as they are important for flood frequency magnitude estimation. We draw the following conclusions from the analysis presented in this paper.

1. The choice of error metric for model calibration impacts high-flow estimates very similarly for both models, although mHM provides overall better performance than VIC in terms of all metrics evaluated.

2. Calibration with KGE improves performance as assessed by high-flow metrics by improving time-dependent metrics (e.g., variability error score). Adjustment of the scaling factors related to the different KGE components (bias, variability, and correlation terms) can further assist the model simulations in matching certain aspects of flow characteristics. The degree of improvement is, however, model dependent.

3. Application-specific metrics can improve estimation of specifically targeted aspects of the system response (here annual peak flows) if used to direct model calibration. However, the use of an application-specific metric does not guarantee acceptable performance with regard to other metrics, even those closely related to the application-specific metric.

Given that Gupta et al. (2009) show clear improvement of flow variability estimates by switching the calibration metric from NSE to KGE for a simple rainfall–runoff model similar to the HBV model (Bergström, 1995), and that our results are similar for two relatively more complex models, we can expect that other models would exhibit similar results when using KGE or its scaled variant. When choosing to use an application-specific metric, it seems clear that careful thought needs to be given to the design of the metric if we are to obtain good performance for both the target metric (used for calibration) and other related metrics (used for evaluation). This is important since we wish to increase confidence in the robustness and transferability of the calibrated model – an issue that needs to be examined in more detail.

Code and data availability. Model calibration was performed using MPR-flex available at https://github.com/NCAR/mpr-flex/tree/direct_calib (last access: 23 August 2017) for VIC 4.1.2h. The mHM 5.8 (https://doi.org/10.5281/zenodo.1069203, Samaniego et al., 2017) is calibrated with the MPR implemented in the model. Hydrometeorological data are obtained from a part of Catchment Attributes and Meteor (CAMELS; Newman et al., 2014; Addor et al., 2017a). Analysis and plotting codes are available at https://github.com/nmizukami/calib4ffa (last access: 15 February 2019).

Author contributions. Authors from NCAR (NM, MPC, AJN, and AWW) and authors from UFZ (OR and RK) initiated model experiment designs separately, and both groups agreed to merge the results. NM, OR and RK performed the model simulations and designed figures and the structure of the paper. HVG provided insights into the model calibration results. All the authors discussed the results and wrote and reviewed the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We thank two anonymous referees for their constructive comments and John Ding for his short comment on NSE. The comments helped improve the manuscript, in particular discussion regarding the consideration of deterministic model residuals for error metric estimates. We also thank Ethan Gutmann and Manabendra Saharia (NCAR) for the earlier discussions on the topic.

Review statement. This paper was edited by Dimitri Solomatine and reviewed by two anonymous referees.

References

Addor, N., Newman, A., Mizukami, N., and Clark, M.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, https://doi.org/10.5065/D6G73C3Q, 2017a.

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

Berghuijs, W. R., Woods, R. A., Hutton, C. J., and Sivapalan, M.: Dominant flood generating mechanisms across the United States, Geophys. Res. Lett., 43, 4382–4390, https://doi.org/10.1002/2016GL068070, 2016.

Bergström, S.: The HBV model, in: Compute Models of Watershed Hydrology, edited by: Singh, V., chap. The HBV mo, Water Resouces Publications, Highlands Ranch Co., 1995.

Bourgin, F., Andréassian, V., Perrin, C., and Oudin, L.: Transferring global uncertainty estimates from gauged to ungauged catchments, Hydrol. Earth Syst. Sci., 19, 2535–2546, https://doi.org/10.5194/hess-19-2535-2015, 2015.

Cuntz, M., Mai, J., Samaniego, L., Clark, M., Wulfmeyer, V., Branch, O., Attinger, S., and Thober, S.: The impact of standard and hard-coded parameters on the hydrologic fluxes in the Noah-MP land surface model, J. Geophys. Res., 121, 10676–10700, https://doi.org/10.1002/2016JD025097, 2016.

de Boer-Euser, T., Bouaziz, L., De Niel, J., Brauer, C., Dewals, B., Drogue, G., Fenicia, F., Grelier, B., Nossent, J., Pereira, F., Savenije, H., Thirel, G., and Willems, P.: Looking beyond general metrics for model comparison – lessons from an international model intercomparison study, Hydrol. Earth Syst. Sci., 21, 423–440, https://doi.org/10.5194/hess-21-423-2017, 2017.

Dieter, F. H. and Arns, S. A.: Estimating Instantaneous Peak Flow from Mean Daily Flow Data, J. Hydrol. Eng., 8, 365–369, https://doi.org/10.1061/(ASCE)1084-0699(2003)8:6(365), 2003.

Ding, J., Wallner, M., Müller, H., and Haberlandt, U.: Estimation of instantaneous peak flows from maximum mean daily flows using

the HBV hydrological model, Hydrol. Process., 30, 1431–1448, https://doi.org/10.1002/hyp.10725, 2016.

Elsner, M., Cuo, L., Voisin, N., Deems, J., Hamlet, A., Vano, J., Mickelson, K. B., Lee, S.-Y., and Lettenmaier, D.: Implications of 21st century climate change for the hydrology of Washington State, Climatic Change, 102, 225–260, https://doi.org/10.1007/s10584-010-9855-0, 2010.

Elsner, M. M., Gangopadhyay, S., Pruitt, T., Brekke, L. D., Mizukami, N., and Clark, M. P.: How Does the Choice of Distributed Meteorological Data Affect Hydrologic Model Calibration and Streamflow Simulations?, J. Hydrometeorol., 15, 1384–1403, https://doi.org/10.1175/jhm-d-13-083.1, 2014.

Farmer, W. H. and Vogel, R. M.: On the deterministic and stochastic use of hydrologic models, Water Resour. Res., 52, 5619–5633, https://doi.org/10.1002/2016WR019129, 2016.

Garcia, F., Folton, N., and Oudin, L.: Which objective function to calibrate rainfall–runoff models for low-flow index simulations?, Hydrolog. Sci. J., 62, 1149–1166, https://doi.org/10.1080/02626667.2017.1308511, 2017.

Gupta, H., Beven, K. J., and Wagener, T.: Model Calibration and Uncertainty Estimation, in: Encyclopedia of Hydrological Sciences, John Wiley & Sons, Ltd, https://doi.org/10.1002/0470848944.hsa138, 2006.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resour. Res., 34, 751–763, https://doi.org/10.1029/97wr03495, 1998.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrol. Process., 22, 3802–3813, https://doi.org/10.1002/hyp.6989, 2008.

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, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., and Andréassian, V.: Large-sample hydrology: a need to balance depth with breadth, Hydrol. Earth Syst. Sci., 18, 463–477, https://doi.org/10.5194/hess-18-463-2014, 2014.

Kavetski, D., Fenicia, F., Reichert, P., and Albert, C.: Signature-Domain Calibration of Hydrological Models Using Approximate Bayesian Computation: Theory and Comparison to Existing Applications, Water Resour. Res., 54, 4059–4083, https://doi.org/10.1002/2017WR020528, 2018.

Klemes, V.: Operational testing of hydrological simulation models, Hydrolog. Sci. J., 31, 13–24, https://doi.org/10.1080/02626668609491024, 1986.

Kumar, R., Samaniego, L., and Attinger, S.: The effects of spatial discretization and model parameterization on the prediction of extreme runoff characteristics, J. Hydrol., 392, 54–69, https://doi.org/10.1016/j.jhydrol.2010.07.047, 2010.

Kumar, R., Livneh, B., and Samaniego, L.: Toward computationally efficient large-scale hydrologic predictions with a multiscale regionalization scheme, Water Resour. Res., 49, 5700–5714, https://doi.org/10.1002/wrcr.20431, 2013a.

Kumar, R., Samaniego, L., and Attinger, S.: Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations, Water Resour. Res., 49, 360–379, https://doi.org/10.1029/2012wr012195, 2013b.

Liang, X., Lettenmaier, D. P., Wood, E. F., and Burges, S. J.: A simple hydrologically based model of land surface water and energy fluxes for general circulation models, J. Geophys. Res., 99, 14415–14428, https://doi.org/10.1029/94jd00483, 1994.

Maurer, E. P., Wood, A. W., Adam, J. C., Lettenmaier, D. P., and Nijssen, B.: A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States, J. Climate, 15, 3237–3251, https://doi.org/10.1175/1520-0442(2002)015<3237:althbd>2.0.co;2, 2002.

Mendoza, P. A., Clark, M. P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G., and Gupta, H.: Are we unnecessarily constraining the agility of complex process-based models?, Water Resour. Res., 51, 716–728, https://doi.org/10.1002/2014WR015820, 2015.

Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijssen, B., Rakovec, O., and Samaniego, L.: Towards seamless large-domain parameter estimation for hydrologic models, Water Resour. Res., 53, 8020–8040, https://doi.org/10.1002/2017WR020401, 2017.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

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

Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., and Nearing, G.: Benchmarking of a Physically Based Hydrologic Model, J. Hydrometeorol., 18, 2215–2225, https://doi.org/10.1175/JHM-D-16-0284.1, 2017.

Olden, J. D. and Poff, N. L.: Redundancy and the choice of hydrologic indices for characterizing streamflow regimes, River Res. Appl., 19, 101–121, https://doi.org/10.1002/rra.700, 2003.

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., and Michel, C.: Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations, Water Resour. Res., 42, W07410, https://doi.org/10.1029/2005WR004636, 2006.

Price, K., Purucker, S. T., Kraemer, S. R., and Babendreier, J. E.: Tradeoffs among watershed model calibration targets for parameter estimation, Water Resour. Res., 48, W10542, https://doi.org/10.1029/2012WR012005, 2012.

Pushpalatha, R., Perrin, C., Moine, N. L., and Andréassian, V.: A review of efficiency criteria suitable for evaluating low-flow simulations, J. Hydrol., 420–421, 171–182, https://doi.org/10.1016/j.jhydrol.2011.11.055, 2012.

Rakovec, O., Kumar, R., Attinger, S., and Samaniego, L.: Improving the realism of hydrologic model functioning through multivariate parameter estimation, Water Resour. Res., 52, 7779–7792, https://doi.org/10.1002/2016wr019430, 2016a.

Rakovec, O., Kumar, R., Mai, J., Cuntz, M., Thober, S., Zink, M., Attinger, S., Schäfer, D., Schrön, M., and Samaniego, L.: Multiscale and Multivariate Evaluation of Water Fluxes and States over European River Basins, J. Hydrometeorol., 17, 287–307, https://doi.org/10.1175/JHM-D-15-0054.1, 2016b.

Samaniego, L., Kumar, R., and Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, Water Resour. Res., 46, W05523, https://doi.org/10.1029/2008wr007327, 2010.

Samaniego, L., Kumar, R., Mai, J., Zink, M., Thober, S., Cuntz, M., and Attinger, S.: mesoscale Hydrologic Model (Version v5.8), Zenodo, https://doi.org/10.5281/zenodo.1069203, 2017.

Samaniego, L., Thober, S., Kumar, R., Wanders, N., Rakovec, O., Pan, M., Zink, M., Sheffield, J., Wood, E. F., and Marx, A.: Anthropogenic warming exacerbates European soil moisture droughts, Nat. Clim. Change, 8, 421–426, https://doi.org/10.1038/s41558-018-0138-5, 2018.

Seiller, G., Roy, R., and Anctil, F.: Influence of three common calibration metrics on the diagnosis of climate change impacts on water resources, J. Hydrol., 547, 280–295, https://doi.org/10.1016/j.jhydrol.2017.02.004, 2017.

Shafii, M. and Tolson, B. A.: Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives, Water Resour. Res., 51, 3796–3814, https://doi.org/10.1002/2014wr016520, 2015.

Shamir, E., Imam, B., Morin, E., Gupta, H. V., and Sorooshian, S.: The role of hydrograph indices in parameter estimation of rainfall–runoff models, Hydrol. Process., 19, 2187–2207, https://doi.org/10.1002/hyp.5676, 2005.

Thober, S., Kumar, R., Wanders, N., Marx, A., Pan, M., Rakovec, O., Samaniego, L., Sheffield, J., Wood, E. F., and Zink, M.: Multi-model ensemble projections of European river floods and high flows at 1.5, 2, and 3 degrees global warming, Environ. Res. Lett., 13, 14003, https://doi.org/10.1088/1748-9326/aa9e35, 2018.

Tolson, B. and Shoemaker, C.: Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, Water Resour. Res., 43, W01413, https://doi.org/10.1029/2005WR004723, 2007.

Westerberg, I. K. and McMillan, H. K.: Uncertainty in hydrological signatures, Hydrol. Earth Syst. Sci., 19, 3951–3968, https://doi.org/10.5194/hess-19-3951-2015, 2015.

Westerberg, I. K., Guerrero, J.-L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J. E., and Xu, C.-Y.: Calibration of hydrological models using flow-duration curves, Hydrol. Earth Syst. Sci., 15, 2205–2227, https://doi.org/10.5194/hess-15-2205-2011, 2011.

Westerberg, I. K., Wagener, T., Coxon, G., McMillan, H. K., Castellarin, A., Montanari, A., and Freer, J.: Uncertainty in hydrological signatures for gauged and ungauged catchments, Water Resour. Res., 1847–1865, https://doi.org/10.1002/2015wr017635, 2016.

Wobus, C., Gutmann, E., Jones, R., Rissing, M., Mizukami, N., Lorie, M., Mahoney, H., Wood, A. W., Mills, D., and Martinich, J.: Climate change impacts on flood risk and asset damages within mapped 100-year floodplains of the contiguous United States, Nat. Hazards Earth Syst. Sci., 17, 2199–2211, https://doi.org/10.5194/nhess-17-2199-2017, 2017.

Wöhling, T., Samaniego, L., and Kumar, R.: Evaluating multiple performance criteria to calibrate the distributed hydrological model of the upper Neckar catchment, Environ. Earth Sci., 69, 453–468, https://doi.org/10.1007/s12665-013-2306-2, 2013.

Yadav, M., Wagener, T., and Gupta, H.: Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, Adv. Water Resour., 30, 1756–1774, https://doi.org/10.1016/j.advwatres.2007.01.005, 2007.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resour. Res., 44, W09417, https://doi.org/10.1029/2007wr006716, 2008.