AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research®

COMMENTARY
10.1029/2020WR029001

Key Points:

- We provide tools to quantify the sampling uncertainty in the Nash-Sutcliffe Efficiency (NSE) and Kling Gupta Efficiency (KGE) metrics
- Our large-sample analysis demonstrates that there is substantial sampling uncertainty in the estimates of NSE and KGE
- We prescribe further research to improve the estimation interpretation, and use of system-scale performance metrics in hydrologic modeling

Correspondence to:
M. P. Clark,
martyn.clark@usask.ca

Citation:
Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., et al. (2021). The abuse of popular performance metrics in hydrologic modeling. *Water Resources Research*, *57*, e2020WR029001. https://doi.org/10.1029/2020WR029001

Received 8 OCT 2020
Accepted 28 JUL 2021

The Abuse of Popular Performance Metrics in Hydrologic Modeling

Martyn P. Clark[1], Richard M. Vogel[2], Jonathan R. Lamontagne[2], Naoki Mizukami[3], Wouter J. M. Knoben[1], Guoqiang Tang[1], Shervan Gharari[4], Jim E. Freer[1], Paul H. Whitfield[1], Kevin R. Shook[4], and Simon Michael Papalexiou[4]

[1]Centre for Hydrology, University of Saskatchewan, Canmore, AB, Canada, [2]Tufts University, Medford, MA, USA, [3]National Center for Atmospheric Research, Boulder, CO, USA, [4]Centre for Hydrology, University of Saskatchewan, Saskatoon, SK, Canada

Abstract The goal of this commentary is to critically evaluate the use of popular performance metrics in hydrologic modeling. We focus on the Nash-Sutcliffe Efficiency (NSE) and the Kling-Gupta Efficiency (KGE) metrics, which are both widely used in hydrologic research and practice around the world. Our specific objectives are: (a) to provide tools that quantify the sampling uncertainty in popular performance metrics; (b) to quantify sampling uncertainty in popular performance metrics across a large sample of catchments; and (c) to prescribe the further research that is, needed to improve the estimation, interpretation, and use of popular performance metrics in hydrologic modeling. Our large-sample analysis demonstrates that there is substantial sampling uncertainty in the NSE and KGE estimators. This occurs because the probability distribution of squared errors between model simulations and observations has heavy tails, meaning that performance metrics can be heavily influenced by just a few data points. Our results highlight obvious (yet ignored) abuses of performance metrics that contaminate the conclusions of many hydrologic modeling studies: It is essential to quantify the sampling uncertainty in performance metrics when justifying the use of a model for a specific purpose and when comparing the performance of competing models.

1. Introduction

A performance metric summarizes the accuracy of a model. In hydrologic modeling, system-scale performance metrics are typically based on the differences between simulated and observed streamflow at the catchment outlet. The most popular system-scale performance metrics in hydrologic modeling are the Nash Sutcliffe Efficiency ($\widehat{\text{NSE}}$; Nash & Sutcliffe, 1970) and the Kling Gupta Efficiency ($\widehat{\text{KGE}}$; Gupta et al., 2009). System-scale performance metrics are widely used as an objective function in model calibration, to justify the use of a model for a specific purpose, and to compare competing models.

The use of performance metrics is constrained by their substantial sampling uncertainty (Lamontagne et al., 2020; Newman, Clark, Sampson, et al., 2015). Such sampling uncertainty can make it difficult to justify the use of a model for specific applications or to compare competing models. For example, $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ have historically been used to define a "good" model, for example, defined as models with $\widehat{\text{NSE}}$ (or $\widehat{\text{KGE}}$) scores above an arbitrarily defined threshold (e.g., see Beven & Binley, 1992; Moriasi et al., 2015). It is uncommon to consider the sampling uncertainty in system-scale metrics when classifying a model as "good" and justifying its use for a specific application. Similarly, it is uncommon to consider the sampling uncertainty in performance metrics when comparing alternative models or during optimization. Given these limitations, it is possible that the selection of models using these metrics cannot be supported, and their conclusions may be suspect.

The purpose of this commentary is to critically evaluate performance metrics that are habitually used in hydrologic modeling. Our specific objectives are three-fold: (a) provide tools to quantify the sampling uncertainty in performance metrics; (b) quantify the sampling uncertainty in the popular performance metrics across a large sample of catchments; (c) prescribe further research that is needed to improve the estimation, interpretation, and use of performance metrics in hydrologic modeling. Our overall intent is to highlight the

obvious (yet ignored) abuses of system-scale performance metrics that contaminate the conclusions from many hydrologic modeling studies.

The remainder of this paper is organized as follows. Section 2 reviews the development of model performance metrics commonly used in hydrologic modeling. Section 3 introduces the database of existing hydrologic model simulations used in this study. Sections 4 and 5 present the results and discussion. Section 6 summarizes the main conclusions of this study.

2. Review of System-Scale Performance Metrics

We examine both the theoretical properties of the Mean Squared Error (MSE), the NSE, and the KGE, as well as their estimation from actual data. We use standard statistical notation where hats denote the sample estimators of theoretical statistics, that is, $\widehat{\text{MSE}}$, $\widehat{\text{NSE}}$, and $\widehat{\text{KGE}}$ define the sample estimators of the theoretical MSE, NSE and KGE statistics. This distinction is necessary to separate the theoretical properties of performance metrics, which do not depend on data, from their sample estimators, which depend on the characteristics of the data in a given modeling application, such as skewness, coefficient of variation, periodicity, persistence, and outliers (Lamontagne et al., 2020).

The MSE, NSE and KGE statistics can be summarized as follows. The MSE is the single most widely used performance metric in the fields of signal processing (Wang & Bovik, 2009) and statistics in general (see Everitt, 2002). The NSE is simply a normalized variant of the MSE (see Equation 6 below). The development of KGE was motivated by algebraic decompositions of the MSE into bias, variance, and correlation components. KGE is only loosely related to NSE and thus MSE, with a complex relationship between NSE and KGE that depends on several factors. For general cases, the relationship between NSE and KGE depends on the coefficient of variation (CV) of the observations (see Equation A1 or sample-based examples for various values of CV in Figure A1 in Knoben et al., 2019, or Equation 10 in Lamontagne et al., 2020). In the special case of unbiased models, the relationship between NSE and KGE still remains complex (e.g., see Figure 1 and Equation 12 of Lamontagne et al., 2020). Lamontagne et al. (2020, Section 2.2) document the unusual conditions under which NSE and KGE are equivalent.

2.1. Mean Squared Error (MSE)

The MSE is a metric that evaluates the goodness of fit between model simulations and observations (Fisher, 1920). The MSE is defined as

$$\text{MSE} = E\left[\left(X_s - X_o\right)^2\right] \tag{1}$$

where $E[\cdot]$ is the expectation operator, and the random variables X_s and X_o define the time series of the model simulations and observations. Once data are introduced, the MSE metric can be estimated from a sample of n pairs of model simulations and observations:

$$\widehat{\text{MSE}} = \frac{1}{n}\sum_{t=1}^{n}\left(x_{s,t} - x_{o,t}\right)^2 \tag{2}$$

where $x_{s,t}$ and $x_{o,t}$ define the model simulations and observations for time step t. Note that the lower-case values in Equation 2, $x_{s,t}$ and $x_{o,t}$ denote sample realizations from the theoretical random variables X_s and X_o.

The expectation in Equation 1 can be expanded to (e.g., see Lamontagne et al., 2020)

$$\text{MSE} = \left(\mu_s - \mu_o\right)^2 + \left(\sigma_s^2 + \sigma_o^2\right) - 2\sigma_s\sigma_o\rho \tag{3}$$

where μ_s and μ_o denote means of the random variables X_s and X_o, σ_s^2 and σ_o^2 denote the variance of X_s and X_o, and ρ defines the Pearson correlation between X_s and X_o. The expansion in Equation 3 was previously derived by Murphy (1988) using sample estimators of the various terms, rather than their population values.

Equation 3, as defined in Murphy (1988), is algebraically identical to Equation 5 in Gupta et al. (2009). Expanding the squared difference in standard deviation as $(\sigma_s - \sigma_o)^2 = \sigma_s^2 - 2\sigma_s\sigma_o + \sigma_o^2$, then

$$\sigma_s^2 + \sigma_o^2 = (\sigma_s - \sigma_o)^2 + 2\sigma_s\sigma_o \tag{4}$$

and substituting Equation 4 in 3, the MSE metric can be written as

$$\text{MSE} = (\mu_s - \mu_o)^2 + (\sigma_s - \sigma_o)^2 + 2\sigma_s\sigma_o(1 - \rho) \tag{5}$$

Equation 5 provides an algebraic decomposition the MSE that includes the bias in the mean (the first term), the standard deviation (the second term) and the covariance (the third term). Note from Equation 5 that the algebraic decomposition of the MSE is not particularly effective because the second and third terms are not independent of one another (see also Gupta et al., 2009; Mizukami et al., 2019).

2.2. The Nash-Sutcliffe Efficiency (NSE)

The $\widehat{\text{NSE}}$ is an estimator of a standardized skill score that measures the fractional improvement over a benchmark. The theoretical version of NSE is

$$\text{NSE} = 1 - \frac{\text{MSE}}{\sigma_o^2} \tag{6}$$

The algebraic decomposition of the NSE can be derived by making use of the decomposition in Equation 3. Substituting Equation 3 into 6 provides a decomposition of the NSE

$$\text{NSE} = -\left(\frac{\mu_s - \mu_o}{\sigma_o}\right)^2 - \left(\frac{\sigma_s}{\sigma_o}\right)^2 + \frac{2\sigma_s\rho}{\sigma_o} \tag{7}$$

Equation 7 is the estimator version in Murphy (1988), his Equation 11, which is identical to the "new" decomposition of NSE presented by Gupta et al. (2009) in their Equation 4, that is,

$$\text{NSE} = -\beta^2 - \alpha^2 + 2\alpha\rho \tag{8}$$

where $\beta = \left(\mu_s - \mu_o\right)/\sigma_o$ and $\alpha = \sigma_s/\sigma_o$. As in Equation 5, the algebraic decomposition of the $\widehat{\text{NSE}}$ is limited because the variance and correlation terms cannot be separated cleanly.

2.3. The Kling-Gupta Efficiency (KGE)

The KGE metric differs from the NSE metric in that it is not derived from the MSE; KGE is simply the Euclidean distance computed using the coordinates of bias, standard deviation, and correlation (Gupta et al., 2009). The theoretical version of the KGE metric is

$$\text{KGE} = 1 - \sqrt{\left(\beta' - 1\right)^2 + \left(\alpha - 1\right)^2 + \left(\rho - 1\right)^2} \tag{9}$$

where $\beta' = \mu_s/\mu_o$. Note that the definition of β' in Equation 9 is different from the definition of β in Equation 8. The bias terms are related as $\beta = (1 - \beta')/\text{CV}_o$ (Knoben et al., 2019), where $\text{CV}_o = \sigma_o/\mu_o$ is the coefficient of variation in the observations.

3. Data and Methods

3.1. Large-Sample Model Simulations for the CAMELS Catchments

In this study we analyze hydrologic model simulations from a large sample of catchments across the contiguous USA (Figure 1). Our analysis uses existing hydrologic model simulations from the Variable Infiltration Capacity model (VIC version 4.1.2h) applied to the 671 catchments in the CAMELS data set (Catchment Attributes and MEteorology for Large-sample Studies). Mizukami et al. (2019) provide details on the large-sample VIC configuration; Newman, Clark, Sampson, et al. (2015) and Addor et al. (2017) provide details on the hydrometeorological and physiographical characteristics of the CAMELS catchments. The CAMELS catchments are those with minimal human disturbance (i.e., minimal land use changes or disturbances, minimal water withdrawals), and are hence almost exclusively smaller, headwater-type catchments (median basin size of 336 km^2).

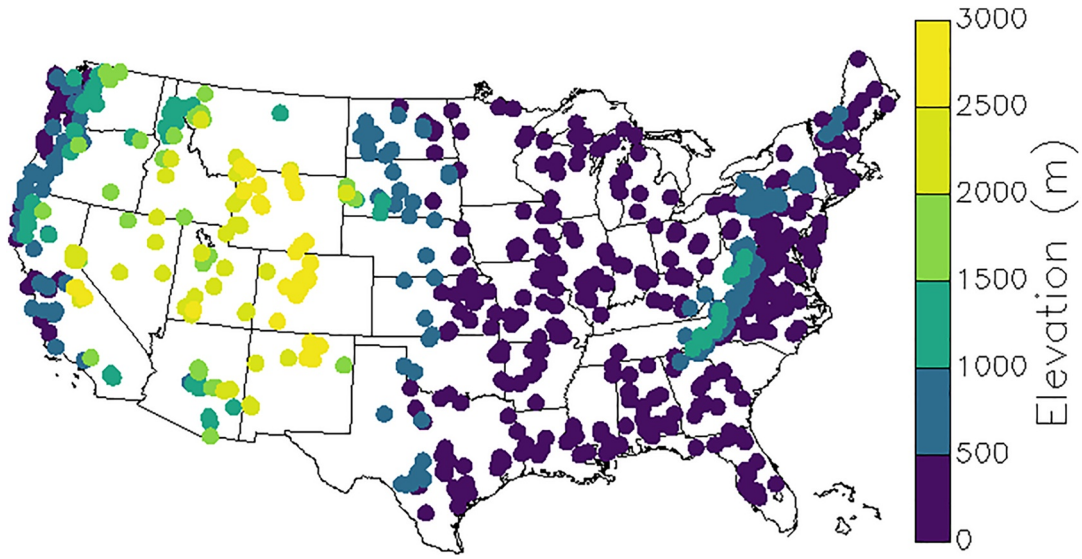

Figure 1. Location and mean elevation of the catchments in the CAMELS data set.

The calibration and evaluation procedure used by Mizukami et al. (2019) is as follows. The VIC model is forced using the daily basin-average meteorological data described by Maurer et al. (2002) and calibrated and evaluated using streamflow data obtained from the USGS National Water Information System server (http://waterdata.usgs.gov/usa/nwis/sw). The VIC model is calibrated using the dynamically dimensioned search (DDS, Tolson & Shoemaker, 2007) algorithm. In each of the 671 CAMELS catchments, the VIC model is calibrated separately for $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ (Mizukami et al., 2019). The hydrometeorological data are split into a calibration period (October 1, 1999–September 30, 2008) and an evaluation period (October 1, 1989–September 30, 1999), with a prior 10-years warm-up period. To maximize the sample size in our analysis, we analyze $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ computed over the combined 19-years calibration and evaluation period (October 1, 1989–September 30, 2008).

3.2. Analysis of the Influence of Individual Data Points

The uncertainties in system-scale performance metrics may be large because the estimates are shaped by a small fraction of the simulation-observation pairs (Clark et al., 2008; Fowler et al., 2018; Lamontagne et al., 2020; McCuen et al., 2006; Newman, Clark, Sampson, et al., 2015; Wright et al., 2019); that is, a small number of simulation-observation pairs have a disproportionate influence on performance metrics. In particular, there is enormous sampling variability associated with streamflow statistics in arid regions (see also Ye et al., 2021). The influence of individual data points can be quantified by successively deleting observations and evaluating their impact on a statistic of interest (e.g., see Efron, 1992; Hampel et al., 1986)—such methods are commonly used in applications of the Jackknife method.

It is straightforward and intuitive to calculate the influence of individual data points on the $\widehat{\text{MSE}}$ estimates. Let $\hat{\varepsilon}_t^2 = (x_{s,t} - x_{o,t})^2$ be the squared difference between simulations $x_{s,t}$ and observations $x_{o,t}$ for a given time step t, and let $\hat{\varepsilon}_{(r)}^2 = \left(\hat{\varepsilon}_{(1)}^2, \hat{\varepsilon}_{(2)}^2, \cdots, \hat{\varepsilon}_{(n)}^2\right)$ be the ranked values of squared errors for all time steps, where $\hat{\varepsilon}_{(1)}^2$ and $\hat{\varepsilon}_{(n)}^2$ are respectively the smallest and largest errors. The influence of the k largest errors on the $\widehat{\text{MSE}}$ estimates, u_k, is simply

$$u_k = \frac{\sum_{i=l}^{n} \hat{\varepsilon}_{(i)}^2}{\sum_{j=l}^{n} \hat{\varepsilon}_{(j)}^2} = \frac{\sum_{i=l}^{n} \hat{\varepsilon}_{(i)}^2}{n\,\widehat{\text{MSE}}} \tag{10}$$

where $l = n - k$. Equation 10 is used in two ways: first, we set $k = 10$ to quantify the influence of the 10 days with the largest errors on the $\widehat{\text{MSE}}$ estimates; second, we identify the k largest observations that contribute to 50% of the $\widehat{\text{MSE}}$ estimates.

3.3. Quantifying Uncertainties in the $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ Estimates

It is particularly important to quantify the sampling uncertainty in model performance metrics when the error distributions exhibit heavy tails, as is the case with the errors obtained from daily streamflow simulations. Parallels to this problem are in the meteorological community, where it is common to quantify the uncertainty in the performance or skill metrics used to describe probabilistic forecasts of rare events (e.g., Bradley et al., 2008; Jolliffe, 2007).

Some attractive approaches to quantify sampling uncertainty are based on the bootstrap (e.g., Vogel & Shallcross, 1996), because they are relatively easy to implement and understand, and because they replace complex theoretical statistical methods with simple brute-force computations (see the Appendix A). Clark and Slater (2006) used bootstrap methods to quantify uncertainties in the performance metrics that they used to evaluate probabilistic estimates of precipitation extremes. Bootstrap methods have also been used to quantify the uncertainty in $\widehat{\text{NSE}}$ estimates (Ritter & Muñoz-Carpena, 2013). Bootstrap methods are likely to find increasing use in hydrology due to the ease with which they can be applied compared to more complex methods. Given their simplicity it is indeed surprising how few examples of the bootstrap there have been in hydrology.

The sampling uncertainty in the $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ estimates is quantified using a mixture of Jackknife and Bootstrap methods. First, we use the Jackknife and Bootstrap methods to compute the standard error in the $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ estimates. These methods resample from the original data sample using the Non-overlapping Block Bootstrap (NBB) strategy of Carlstein (1986), using data blocks of length one year. The use of data blocks of length one year reduces the issues with substantial seasonal non-stationarity in shorter data blocks, while preserving the within-year autocorrelation and seasonal periodicity of streamflow series. Bootstrapping methods are only effective if the blocks used are approximately independent. Second, we use the Bootstrap methods to compute tolerance intervals for the $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ estimates, where the 90% tolerance intervals are defined as the difference between the 95th and 5th percentile of the empirical probability distribution of the $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ estimates. Tolerance intervals differ from confidence intervals, because tolerance intervals are intervals corresponding to a random variable, rather than random confidence intervals around some true value. These bootstrap tolerance intervals are computed using 1,000 bootstrap samples. Finally, we use the Jackknife-After-Bootstrap method (Efron, 1992) to estimate the standard error in the Bootstrap tolerance intervals, which enables us to evaluate how sensitive the resulting uncertainty intervals are to individual years (blocks). The implementation details of the uncertainty quantification methods discussed above are summarized in the Appendix A; the open-source "gumboot" package has been developed to quantify the sampling uncertainty in performance metrics (https://github.com/CH-Earth/gumboot; https://cran.r-project.org/package=gumboot).

It is important to note that the methods implemented here quantify the sampling uncertainty in the $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ estimates for a given hydrologic model and a given sample of streamflow observations. The model itself will contain uncertainty (e.g., uncertainty in the meteorological inputs; uncertainty in the model parameters and model structure). The observations used to compute the $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ estimates also contain uncertainty, especially for the high flow extremes that can have a large influence on the $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ estimates. The model and data uncertainty are not explicitly included in the estimates of sampling uncertainty (we will return to this point in Section 5.3).

4. Results

The probability distribution of squared errors between model simulations and observations have heavy tails, meaning that the estimates of sum-of-squared error statistics can be heavily influenced by a small fraction of the simulation-observation pairs (Clark et al., 2008; Fowler et al., 2018; Lamontagne et al., 2020;

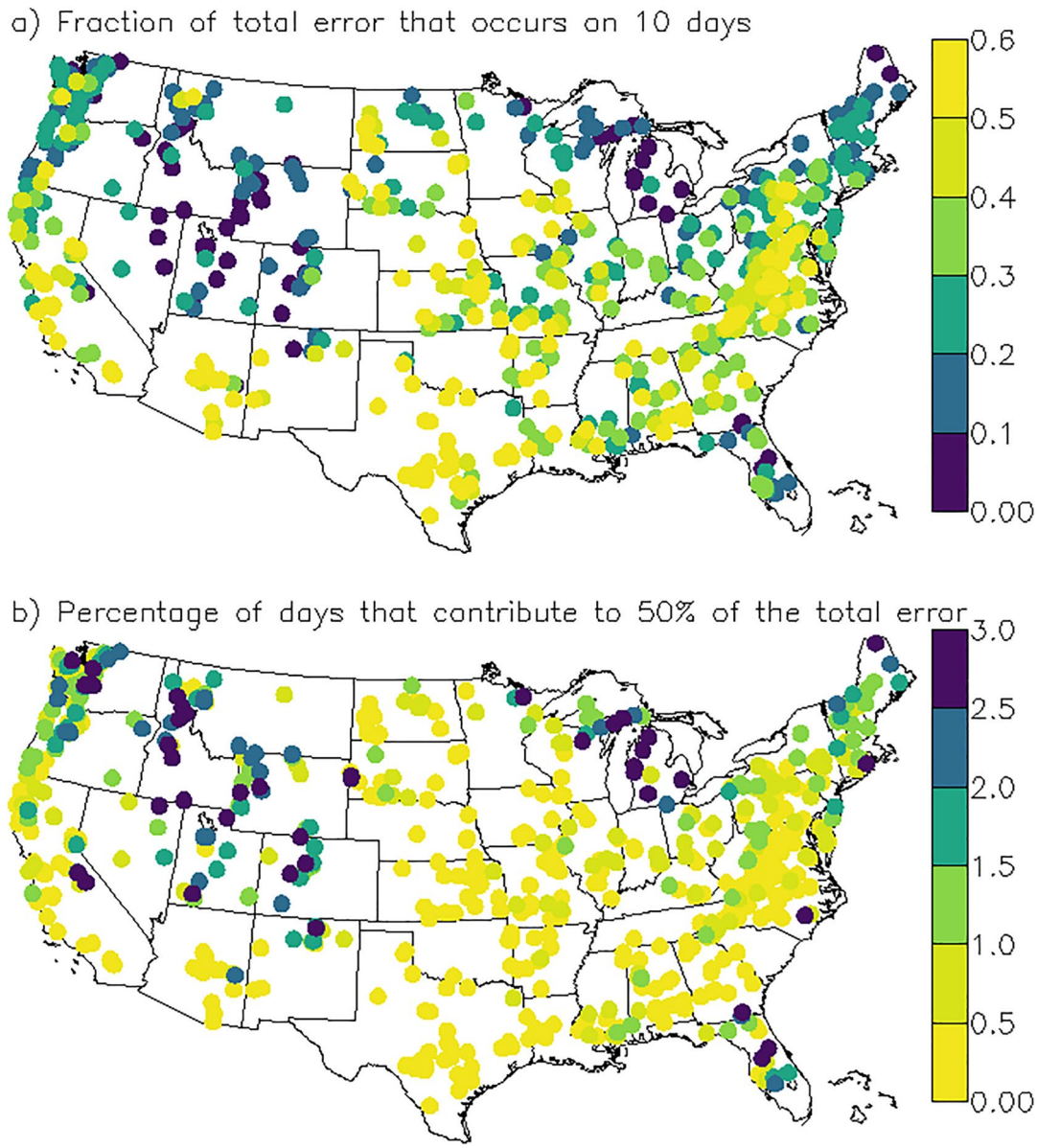

Figure 2. Contribution of subset of days to the $\widehat{\text{MSE}}$ estimate. The upper plot shows the fraction of the $\widehat{\text{MSE}}$ estimate contributed by the 10 days with the highest error. The lower plot shows the percentage of days that contribute to 50% of the $\widehat{\text{MSE}}$ estimate.

Newman, Clark, Sampson, et al., 2015). To document this issue, Figure 2 uses Equation 10 to quantify the influence of the k largest errors on the $\widehat{\text{MSE}}$ estimates, repeating the analysis of Newman, Clark, Sampson, et al. (2015) with the VIC model. Figure 2a quantifies the influence of the 10 individual days with the largest errors on the $\widehat{\text{MSE}}$ estimates–Figure 2a demonstrates that, in many catchments, 10 days in the 19-year period contribute to over 50% of the sum-of-squared errors between simulated and observed streamflow. Figure 2b identifies the k largest observations that jointly contribute 50% of the $\widehat{\text{MSE}}$ estimate, expressed as a percentage of the total sample length n. Figure 2b demonstrates that, in many catchments, 50% of the sum-of-squared errors is caused by less than 0.5% of the simulation-observation pairs. These results suggest that there will be large uncertainty in the $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ metrics.

Figure 3 quantifies the uncertainty in $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ across the CAMELS catchments, illustrating considerable uncertainty in both the $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ values. Figure 3 illustrates that the 90% tolerance intervals for

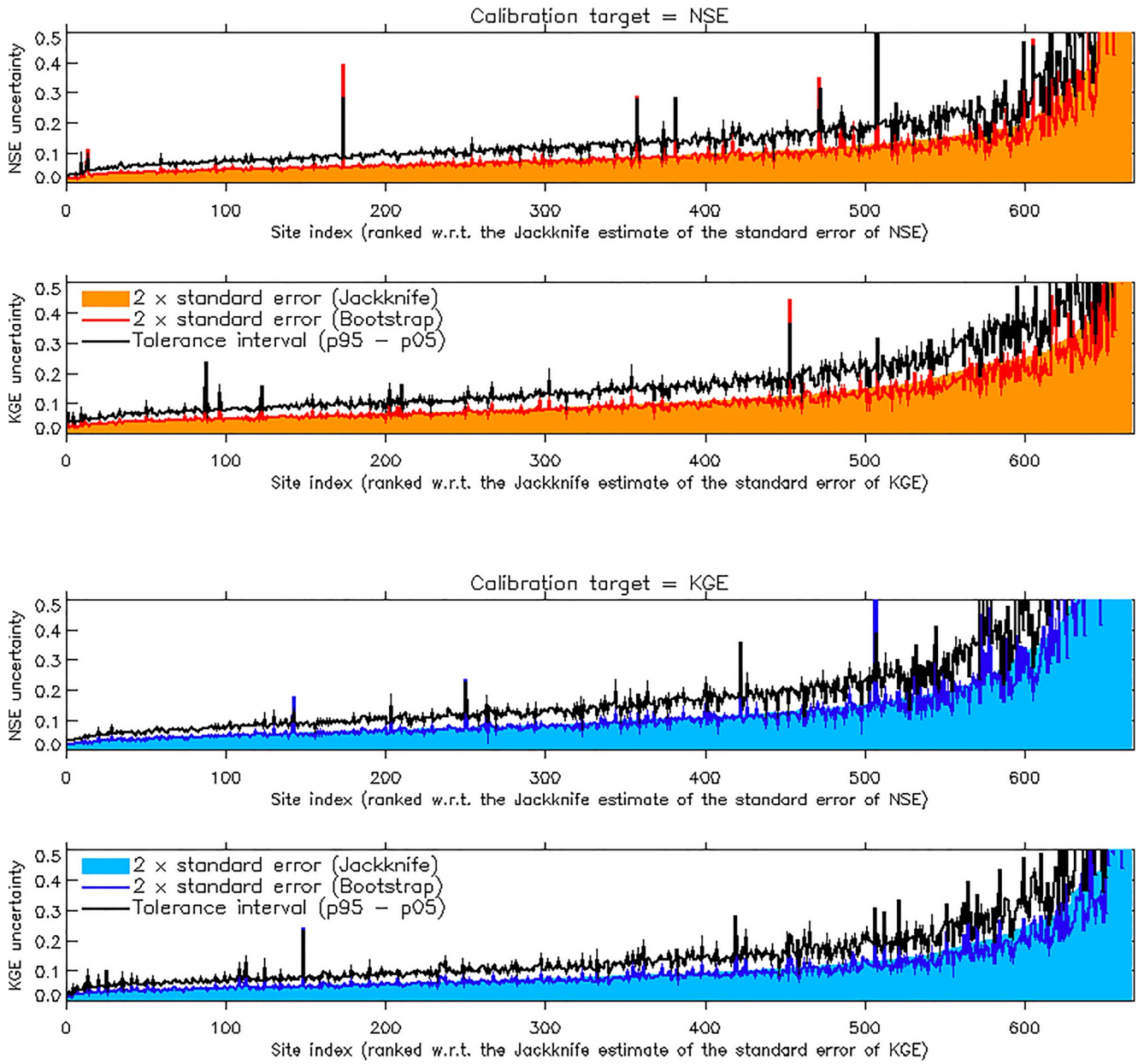

Figure 3. Estimates of uncertainty in the $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ estimates across the CAMELS catchments. The uncertainty is quantified using standard error estimates (×2) obtained using Jackknife and Bootstrap estimates (see the Appendix A for implementation details), along with tolerance intervals computed as the difference between the 95th and 5th percentiles of the Bootstrap samples. Results are shown for calibrations obtained by maximizing the $\widehat{\text{NSE}}$ metric (upper plots) and by maximizing the $\widehat{\text{KGE}}$ metric (lower plots).

both NSE and KGE (as obtained by the bootstrap methods described in the Appendix A) are greater than 0.1 for more than half of the CAMELS catchments. The results in Figure 3 illustrate that both the bootstrap and jackknife methods yield consistent standard error estimates. The large uncertainty in $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ are evident when both $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ are used as a calibration target.

The jackknife-after-bootstrap methods enable an evaluation of the degree of precision and accuracy associated with the bootstrap tolerance intervals. While there is considerable sampling uncertainty in the tolerance intervals (estimated using the jackknife-after-bootstrap methods; Figure 4), that uncertainty is considerably smaller than the uncertainty associated with $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ as is shown in Figure 3. As we dis-

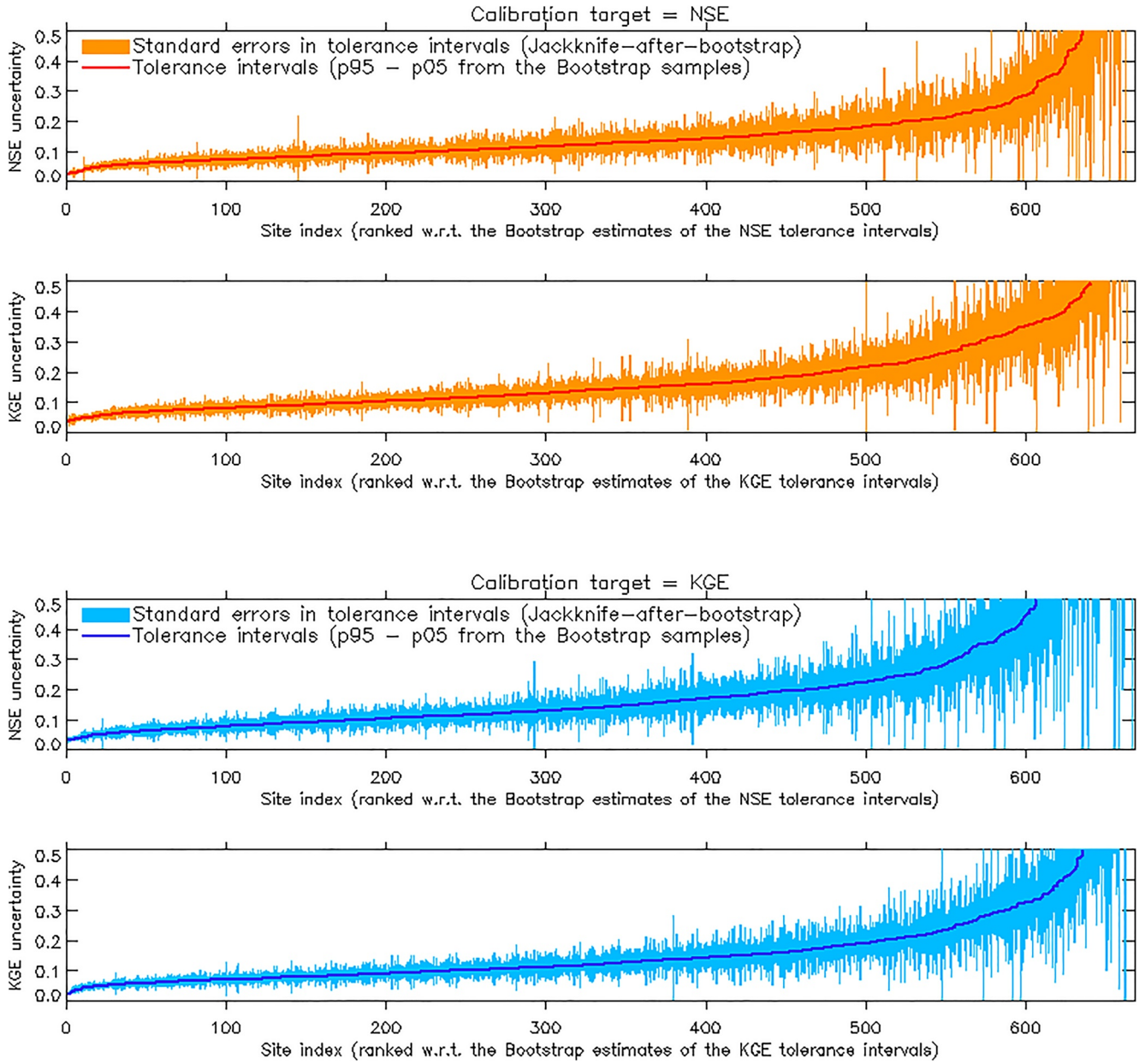

Figure 4. Standard error in the Bootstrap tolerance intervals shown in Figure 3. The standard error in the Bootstrap tolerance intervals is estimated using the jackknife-after-bootstrap method of Efron (1992), as summarized in the Appendix A. Results are shown for calibrations obtained by maximizing the $\widehat{\text{NSE}}$ metric (upper plots) and by maximizing the $\widehat{\text{KGE}}$ metric (lower plots).

cuss in the next section, the sampling uncertainty depicted in Figure 3 may be under-estimated in situations where there is extremely high skewness in daily streamflows.

5. Discussion

5.1. It Is Necessary to Quantify the Uncertainty in Performance Metrics

The high uncertainty associated with the estimators $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ underscores the need to quantify the uncertainty in the performance metric estimators used in hydrologic modeling applications. Quantifying the sampling uncertainty in model evaluation statistics is easily accomplished using appropriate bootstrap methods. Moreover, bootstrap methods can be applied to any performance metric estimator. Quantifying

the uncertainty in the performance metric estimators should arguably become a routine part of the hydrologic modeling enterprise. As our results show, the width of the 90% tolerance intervals associated with the estimators $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ are greater than 0.1 in at least half of analyzed catchments. Such wide 90% tolerance intervals indicate considerable uncertainty associated with each of these metrics. These results imply that the conclusions from many hydrologic modeling studies may not be justified in light of the high sampling uncertainty in system-scale performance metric estimators.

In spite of the ease with which the bootstrap may be applied as a post-processing approach to developing uncertainty intervals, there is a need for additional research on methods to quantify the sampling uncertainty. Our experiments (not shown) demonstrate that traditional bootstrap methods may severely under-estimate the sampling uncertainty in the estimators $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ in situations where there is extremely high skewness (see also Chernick & LaBudde, 2011). These under-estimates in uncertainty occur because bootstrap methods "recycle" the observations, and the bootstrap samples do not adequately encapsulate the uncertainty associated with the few extraordinary errors in the thick upper tail of the error distribution. Indeed, our Jackknife-after-Bootstrap analyses demonstrate that there are large standard errors in our bootstrap estimates of uncertainty in $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$. Thus, given the extremely high skewness of daily streamflow observations in some watersheds, we recommend future research which compares the uncertainty intervals derived from various bootstrap methods against the uncertainty intervals derived from more advanced stochastic methods (e.g., Papalexiou, 2018).

5.2. It Is Necessary to Improve the Estimates of System-Scale Performance Statistics

A variety of approaches can be introduced to improve estimates of the theoretical NSE and KGE statistics; that is, to develop more robust estimates $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ that have lower sampling uncertainty. For example, Fowler et al. (2018) calculated the $\widehat{\text{KGE}}$ metric separately for each year before averaging across years; Lamontagne et al. (2020) introduced alternative estimators of $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ based on a bivariate lognormal monthly mixture model. Variance reduction methods introduced to the field of machine learning and statistics (e.g., Nelson & Schmeiser, 1986) can be used to improve estimates of the theoretical NSE and KGE performance metrics. More generally, the approaches of bagging and bragging could be tested, where the performance metrics are estimated using the median or the mean of multiple bootstrap samples (Berrendero, 2007). Further work is needed to better understand the characteristics of data points that have high leverage in order to devise methods that improve estimates of the theoretical NSE and KGE statistics.

5.3. It Is Necessary to Put Performance Metrics in Context

The growing field of model benchmarking seeks to put performance metrics into context, for example, by asking the question if models meet our *a-priori* expectations, or if models adequately use the information that is available to them. The recent efforts in model benchmarking have focused on defining lower and upper benchmarks to provide context for model performance (Nearing et al., 2018; Newman et al., 2017; Seibert et al., 2018). Lower benchmarks evaluate the extent to which models surpass expectations (Seibert, 2001), for example, the extent to which model simulations perform better than a benchmark such as climatology, persistence, simulations from another model (Wilks, 2011), or departures from the seasonal cycle (Knoben et al., 2020; Schaefli & Gupta, 2007). A key component of defining the lower benchmark is defining our *a-priori* expectations of model capabilities. We define the upper benchmark to quantify the predictability of the system, that is, the maximum information content in the forcing-response data (Nearing et al., 2018; Newman et al., 2017). For example, Best et al. (2015) recently demonstrated that many mechanistic land models were out-performed by simple statistical models, implying that modern land models were not adequately using the information that is available to them. Much work still needs to be done to quantify our expectations for model performance (the lower benchmark) as well as to quantify system-scale predictability (the upper benchmark).

Benchmarking is important in the context of performance metrics because the NSE and KGE have rather weak *a-priori* expectations of model performance. The NSE uses the variance of the observations as the benchmark. This means that NSE > 0 if the MSE is smaller than the variance of the observations. In other words, NSE > 0 if the model simulations are better than the reference case where $x_{s,t} = \mu_o$ for all time steps.

Knoben et al. (2019) points out that the KGE estimates do not have the same benchmark as NSE estimates: the implied benchmark associated with estimates of NSE, that is, that model simulations are always equal to the observed mean (i.e., NSE = 0) occurs when the estimate of KGE = $1 - \sqrt{2}$ (i.e., when the estimate of KGE = –0.41). The observed mean is often used as a benchmark with the KGE metric as well, imposing old expectations on a new metric. Using stricter, purpose-specific benchmarks can give a clearer idea of model strengths and weaknesses.

It is also necessary to evaluate the system-scale performance metrics in the context of the uncertainties in the model inputs (e.g., spatial meteorological forcing data), the uncertainties in the hydrologic model (e.g., uncertainties in model parameters and model structure), and the uncertainties in the system-scale response (e.g., streamflow observations). Many groups are now developing ensemble spatial meteorological forcing fields in order to understand how uncertainties in the model forcing data affect uncertainties in the hydrologic model simulations (e.g., Clark & Slater, 2006; Cornes et al., 2018; Frei & Isotta, 2019; Newman, Clark, Craig, et al., 2015; Tang, Clark, Papalexiou, et al., 2021). There are also a wealth of approaches to quantify hydrologic model uncertainty. Vogel (2017) introduced the concept of stochastic watershed models (SWM), which involve methods for generating likely stochastic traces of daily streamflow from deterministic watershed models. All such methods of developing SWMs reviewed by Vogel (2017), including the very generalized blueprint introduced by Montanari and Koutsoyiannis (2012), may be employed to develop uncertainty intervals associated with either streamflow predictions or other water resource system variables. There is now also substantial effort dedicated to quantifying uncertainty in streamflow observations (e.g., see the comparison of uncertainty techniques by Kiang et al., 2018 and also Coxon et al., 2015 and Mansanarez et al., 2019). The key issue is that the most uncertain observations of streamflow are in the upper tail; these observations also have the most influence on the KGE and NSE metrics. Further research is needed to understand how these sources of uncertainty are manifest in system-scale performance metrics.

5.4. It Is Necessary to Understand the Limitations of System-Scale Performance Metrics

It is well known that minimizing the sum-of-squared errors in calibration results in simulated streamflows with smaller variance than the observations (e.g., Gupta et al., 2009). This occurs because of the interplay between estimates of the variance of the flows and correlation in NSE described in Section 2.2–specifically, the quantity α appears in both the second and third terms in Equation 8, meaning that NSE is maximized when $\alpha = \rho$. This is problematic because optimization studies that minimize the $\widehat{\text{MSE}}$ (or maximize the $\widehat{\text{NSE}}$) result in $\hat{\alpha} < 1$ because $\hat{\rho}$ is always smaller than unity. Mizukami et al. (2019) illustrate these issues when using $\widehat{\text{NSE}}$ as an objective function in large-sample hydrologic model calibration study. They showed that the calibrated simulations had substantial under-estimates of high flow events, such as the annual peak flows that are used for flood frequency estimation. Underestimation of variance, as well as all other upper moments, is a general problem associated with simulation models and is not limited to use of a particular objective function (see Farmer & Vogel, 2016).

There are also problems with the KGE metric. As discussed by Santos et al. (2018), the definition of the bias term in KGE, $\beta' = \mu_s / \mu_o$, can lead to very large values of β' (and hence low KGE scores) when μ_o is small. Such problems with amplified β' values are potentially more pronounced for variables where μ_o crosses zero (e.g., log-transformed flows, temperature) because μ_o could be very small. Citing drawbacks of the NSE as justification, part of the community has switched to using $\widehat{\text{KGE}}$ over $\widehat{\text{NSE}}$. We argue that this did not solve but only changed the problems related to system-scale performance metrics. It is important to be aware of the theoretical behavior of system-scale performance metrics, along with their limits of applicability, and use additional metrics that are tailored to suit specific applications.

5.5. It Is Necessary to Use Additional Performance Metrics

A key problem with system-scale performance metrics is that they do not make adequate use of the full information content in the data. Gupta et al. (2008) point out that global calibration of hydrologic models (e.g., using $\widehat{\text{NSE}}$ or $\widehat{\text{KGE}}$ as the objective function) entails compressing the information in the model output and observations into a single performance metric, and then using that single metric to infer values of multiple model parameters and all aspects of hydrological processes. Such global calibration methods can

lead to problems of compensatory parameters, providing the "right" results for the wrong reasons (Kirchner, 2006). Specifically, parameters in one part of the model may be assigned unrealistic values that compensate for unrealistic parameter values in another part of the model, or that compensate for errors in the model forcing data and weaknesses in model structure (Clark & Vrugt, 2006). Addressing this problem requires asking a different question: Instead of asking "how good is my model?", it may be more appropriate to ask "What is my model good for?" This second question is more relevant when designing a modeling experiment for a specific application.

One approach is to develop alternative system-scale performance metrics. This includes the efforts to develop variants of KGE–for example, Kling et al. (2012) introduced a modified version of KGE, termed KGE′, by using the ratio of the simulated and observed coefficient of variation (CV $= \sigma/\mu$) instead of the ratio of the simulated and observed standard deviation. Their intent is to reduce the impact of bias on the variability term in KGE. Pool et al. (2018) developed alternative estimates of $\hat{\alpha}$ and $\hat{\rho}$ for use with $\widehat{\text{KGE}}$, with the intent of reducing the impact of outliers. Note that the Pool et al. (2018) estimates of $\hat{\alpha}$ and $\hat{\rho}$ are for different theoretical statistics than the α and ρ statistics that are used in Equations 8 and 9 (see Barber et al., 2019; Lamontagne et al., 2020). Other alternative system-scale metrics include variable transformations, such as the log-transform or Box-Cox transform (to reduce skewness, and focus more on low flows), or methods to compare distributions of modeled extremes to observed extremes. The work to develop alternative system-scale performance metrics recognizes that estimates of correlation-based metrics are often inflated, in the sense that high values can occur for mediocre and poor models, and that estimators of correlation-based metrics are sensitive to outliers and data asymmetry (Legates & McCabe, 1999; Willmott, 1981; see also Mo et al., 2014; Barber et al., 2019).

In this context, it is worth pointing out that it is straightforward to redefine the KGE metric to address the problems with the amplified β' values described above. For example, the bias component of the mean in the KGE metric could be represented as $(\mu_s - \mu_o)^2/{\sigma_o}^2$, as it is in the NSE metric. It is hence straightforward to modify the KGE metric such that

$$\text{KGE}'' = 1 - \sqrt{\beta^2 + (\alpha - 1)^2 + (\rho - 1)^2} \tag{11}$$

where, as in Equation 7, $\beta = (\mu_s - \mu_o)^2/{\sigma_o}^2$. The KGE'' metric has been used by Tang et al. (2021a, 2021b). These modifications to the KGE metric avoid the amplified β' values when μ_o is small. Note that since σ_o is constrained to be positive, the zero-bounded structure of σ_o means that normalizing by σ_o will not have the same problems as normalizing by μ_o in the original KGE or KGE' metrics.

Another approach is to use additional non-global metrics (e.g., multiple diagnostic signatures of hydrologic behavior). For example, much of the research on model calibration and evaluation now focuses on multi-criteria methods, including analysis of trade-offs among multiple objective functions (e.g., Fenicia et al., 2007; Yapo et al., 1998), analysis of the temporal variability of model errors (Coxon et al., 2014; Reusser et al., 2009), and scrutinizing diagnostic signatures of hydrologic behavior in order to identify model weaknesses (Gupta et al., 2008; Rakovec et al., 2016). A key part of this analysis is to understand the sensitivity of different non-global metrics to individual parts of a model (e.g., Markstrom et al., 2016; Van Werkhoven et al., 2009). As such, these alternative metrics can focus attention on aspects of the model that may be more relevant for specific modeling applications.

6. Conclusions

The goal of this commentary is to critically evaluate the performance metrics that are habitually used in hydrologic modeling. Our focus is on the Nash-Sutcliffe Efficiency (NSE) and the Kling-Gupta Efficiency (KGE) metrics, which are both widely used in science and applications communities around the world. Our contributions in this paper are three-fold:

1. We provide tools to enable hydrologic modelers to quantify the sampling uncertainty in system-scale performance metrics. We use the non-overlapping block bootstrap method to obtain probability distributions and associated tolerance intervals of estimates of NSE and KGE, and we use the jackknife-after-bootstrap method to obtain estimates of standard error of those bootstrap tolerance intervals. These

comparisons enable us to ensure that even though the tolerance intervals display sampling variability, that variability is always considerably smaller than the tolerance intervals themselves, thus providing a nice validation of the precision of the tolerance intervals.

2. We quantify the sampling uncertainty in system-scale performance metrics across a large sample of catchments. Our results show that the probability distribution of squared errors between model simulations and observations have heavy tails, meaning that the estimates of sum-of-squared error statistics can be shaped by just a few simulation-observation pairs (Figure 2). This leads to substantial uncertainty in the estimators $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ (Figures 3 and 4). The implication of these results is that the conclusions from many hydrologic modeling studies are based on values for these metrics that fall well within the metrics' uncertainty bounds. Such conclusions may thus not be justified.
3. We define further research that is, needed to improve the estimation, interpretation, and use of system-scale performance metrics in hydrological modeling

More generally, our commentary highlights the obvious (yet ignored) abuses of performance metrics that contaminate the conclusions of many hydrologic modeling studies. We look forward to additional studies that improve the scientific basis of model evaluation.

Appendix A: The Jackknife and Bootstrap Methods

In this study, we use two resampling methods, the Jackknife and the Bootstrap, to estimate the empirical probability distribution of the $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ estimators for each of the 671 CAMELS catchments. These methods estimate the empirical probability distribution of a given statistic by drawing or resampling a number of independent samples from the original sample of data.

The following sub-sections describe the implementation of the Jackknife and Bootstrap methods, including the resampling strategies, the Jackknife and Bootstrap estimates of standard error, and the Jackknife estimates of the standard error in the bootstrap-derived empirical probability distributions of $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$.

A1. The Jackknife and Bootstrap Resampling Strategies

The Jackknife method is a structured approach of resampling without replacement where observations are successively deleted from the original sample of data. A Jackknife sample is the data set that remains after deleting the ith observation, or deleting the ith block of observations, that is,

$$x_{(i)} = \left(x_1, \&, x_{i-1}, x_{i+1}, \&, x_n\right) \tag{A1}$$

The value of the ith Jackknife replicate is the value of the estimator $\hat{\theta}_{(i)} = g\left(x_{(i)}\right)$. In our case, the ith Jackknife replicate is $\hat{\theta}_{(i)} = \widehat{\text{NSE}}_{(i)}$ or $\hat{\theta}_{(i)} = \widehat{\text{KGE}}_{(i)}$. The Jackknife method is useful in cases where it is desirable to conduct structured analysis of the deleted point statistics.

The Bootstrap method is much more flexible than the Jackknife method. The Bootstrap method uses the approach of resampling with replacement. A Bootstrap sample is obtained by using a random number generator to make n independent draws from the original sample of data (Efron & Tibshirani, 1986), that is,

$$y = \left(x_1^*, x_2^*, \cdots, x_n^*\right) \tag{A2}$$

and the process is repeated to generate B samples, that is, $y_{(1)}^*, y_{(2)}^*, \cdots, y_{(B)}^*$. Then, for each sample $y_{(b)}^*$ compute the statistic of interest, that is, $\hat{\theta}_{(b)}^* = g\left(y_{(b)}^*\right), b = 1, 2, \cdots, B$. The empirical probability distribution of the statistic of interest can then be calculated using all of the B samples.

When implementing these resampling methods, it is necessary to ensure independence between each draw from the original sample of data (Carlstein, 1986; Künsch, 1989; Vogel & Shallcross, 1996). Specifically, the errors in daily streamflow simulations are characterized by substantial periodicity and persistence – this creates complex temporal dependence structures on time scales from days (e.g., errors in the simulations of recessions after a storm event) to seasons (e.g., errors in the simulations of seasonal snow accumulation and

melt, or errors in the seasonal cycle of transpiration). To address these issues, we implement a non-overlapping block resampling strategy that was developed for the Bootstrap method, the Non-overlapping Block Bootstrap (NBB) of Carlstein (1986). This approach identifies k subseries of data of length λ, where each sub-series of data is statistically independent. In our implementation, the k subseries are each of the 19 water years $(1990,1991,\cdots,2008)$, where the water years span the period Oct 1st–Sep 30th (e.g., water year 1990 is the period Oct 1st1989–Sep 30th 1990).

The non-overlapping block resampling strategy is used for both the Jackknife and Bootstrap methods. The Jackknife sample for a given water year is the data set that remains after deleting the ith water year. For example, $x_{(2002)} = (x_{1990},\ldots,x_{2001},x_{2003},\ldots,x_{2008})$, where $x_{(2002)}$ contains the daily time series of simulation-observation pairs for all years except water year 2002, and $\hat{\theta}_{(2002)} = g(x_{(2002)})$ is the $\widehat{\text{NSE}}$ or $\widehat{\text{KGE}}$ estimates using all daily data except in 2002. The Bootstrap method samples water years with replacement: A given bootstrap sample may include a given water year more than once, or a given sample may not include a given water year at all. The Bootstrap samples that do not have a given water years (e.g., all Bootstrap samples without water year 2002) open up opportunities to quantify the standard errors in the Bootstrap estimates of the empirical probability distributions (using the Jackknife-After-Bootstrap method introduced by Efron, 1992; we will discuss this implementation in Section A3).

A2. Jackknife and Bootstrap Estimates of Standard Errors in the $\widehat{\text{NSE}}$ and $\widehat{\text{KGE}}$ Estimates

The Jackknife estimates of standard error can be obtained by first considering the case where the standard error estimates are not needed (Efron & Gong, 1983). The average of the jackknife sample, $\bar{x}_{(i)}$, is

$$\bar{x}_{(i)} = \frac{\left(\sum_{j=1}^{n} x_j\right) - x_i}{n-1} = \frac{n\bar{x} - x_i}{n-1} \tag{A3}$$

with the ith observation given as $x_i = n\bar{x} - (n-1)\bar{x}_{(i)}$. The standard error of $\bar{x}$ is then (Efron & Gong, 1983)

$$\widehat{\text{se}}_{\text{jack}}(\bar{x}) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\bar{x}_{(\cdot)} - \bar{x}_{(i)}\right)^2} \tag{A4}$$

with $\bar{x}_{(\cdot)} = \sum_{i=1}^{n} \bar{x}_{(i)}/n$.

Equation A4 can be extended to compute the standard error for any statistic of interest. If we let $\hat{\theta}_{(i)} = g(x_{(i)})$ be the deleted point value for a given statistic (Efron, 1992), then the jackknife estimate of the statistic of interest, $\hat{\theta}_{\text{jack}}$, can be defined as

$$\hat{\theta}_{\text{jack}} = n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)} \tag{A5}$$

where $\hat{\theta}$ is the estimate of the statistic using all observations and $\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \hat{\theta}_{(i)}/n$. The standard error of $\hat{\theta}_{\text{jack}}$ is then

$$\widehat{\text{se}}_{\text{jack}}(\hat{\theta}) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)}\right)^2} \tag{A6}$$

The Bootstrap estimate of standard error is more straightforward: It is simply the standard deviation of the Bootstrap samples, that is,

$$\widehat{\text{se}}_{\text{boot}}(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^*_{(b)} - \hat{\theta}^*_{(\cdot)}\right)^2} \tag{A7}$$

where $\hat{\theta}^*_{(\cdot)} = \sum_{b=1}^{B} \hat{\theta}^*_{(b)}/B$.

A3. Jackknife Estimates of Standard Error in the Bootstrap-Derived Probability Distributions

The Bootstrap estimates of the empirical probability distributions create a conundrum: whilst outliers can cause large uncertainty in the $\widehat{\text{NSE}}$ or $\widehat{\text{KGE}}$ estimates, the outliers can also create large uncertainty in the Bootstrap estimates of the empirical probability distributions. It is hence necessary to estimate the standard error in the Bootstrap methods.

Estimates of the standard error in the Bootstrap methods can be computed easily using the Jackknife-After-Bootstrap method of Efron (1992). In the previous discussion we noted that the non-overlapping block resampling strategy opens up opportunities to quantify the standard errors in the Bootstrap estimates of the empirical probability distributions. Specifically, for a given water year we can compute a Jackknife sample using all of the Bootstrap samples that do not include that water year. When such Jackknife samples are constructed for all water years, the Jackknife method can be used to estimate standard error in the Bootstrap estimates (the Jackknife-After-Bootstrap method).

The Jackknife-After-Bootstrap method is implemented as follows (Efron, 1992). Our starting point is the B estimates of the statistic of interest, that is, $\hat{\theta}^* = \left(\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \cdots, \hat{\theta}^*_{(B)}\right)$, that were computed from the B samples $y^*_{(1)}, y^*_{(2)}, \cdots, y^*_{(B)}$ obtained from the Bootstrapping. Recall that each of the B samples is constructed by making n draws from the original sample of data, that is, $y = \left(x^*_1, x^*_2, \cdots, x^*_n\right)$. Given this information, we can calculate the proportion of each Bootstrap sample that equals a given observation x_i, that is, (Efron, 1992),

$$P^*_{i,b} = \#\left\{y^*_{(b)} = x_i\right\}/n,\ b = 1,2,\cdots,B \tag{A8}$$

and define the resampling vector for a given observation,

$$P^*_i = \left(P^*_{i,1}, P^*_{i,2}, \cdots, P^*_{i,B}\right) \tag{A9}$$

It is then straightforward to identify the subset of bootstrap samples where $P^*_i = 0$ (i.e., the subset of bootstrap samples that do not include the observation x_i) and define samples of the statistic of interest where $P^*_i = 0$,

$$\hat{\theta}^*_i = \left\{\hat{\theta}^* | P^*_i = 0\right\} = \left(\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \cdots, \hat{\theta}^*_{(D)}\right) \tag{A10}$$

where $D = \#\left\{P^*_i = 0\right\}$ and $\hat{\theta}^* = \left(\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \cdots, \hat{\theta}^*_{(B)}\right)$ is the statistic of interest for all bootstrap samples. It is then possible to compute statistics from the subset of Bootstrap samples, that is,

$$\hat{\gamma}^*_{(i)} = g\left(\hat{\theta}^*_i\right) \tag{A11}$$

where $g(\cdot)$ may be a statistic such as the fifth or 95th percentile.

The Jackknife estimate of standard error uses Equation A6 with $\hat{\gamma}^*_{(i)}$ as the value of the ith Jackknife replicate in place of $\hat{\theta}_{(i)}$.

Data Availability Statement

The data for the large-domain model simulations are publicly available at the National Center for Atmospheric Research at https://ral.ucar.edu/solutions/products/camels. The source code to quantify the sampling uncertainty in performance metrics (the "gumboot" package) is available at https://github.com/CH-Earth/gumboot.

Acknowledgments
We appreciate the constructive comments from the four reviewers. Martyn Clark, Wouter Knoben, Guoqiang Tang, Shervan Gharari, Jim Freer, Paul Whitfield, Kevin Shook, and Simon Papalexiou were supported by the Global Water Futures program, University of Saskatchewan.

References

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Barber, C., Lamontagne, J., & Vogel, R. M. (2019). Improved estimators of correlation and R^2 for skewed hydrologic data. *Hydrological Sciences Journal*, *65*(1), 87–101. https://doi.org/10.1080/02626667.2019.1686639

Berrendero, J. R. (2007). The bagged median and the bragged mean. *The American Statistician*, *61*(4), 325–330. https://doi.org/10.1198/000313007x245401

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., et al. (2015). The plumbing of land surface models: Benchmarking model performance. *Journal of Hydrometeorology*, *16*(3), 1425–1442. https://doi.org/10.1175/JHM-D-14-0158.1

Beven, K., & Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, *6*(3), 279–298. https://doi.org/10.1002/hyp.3360060305

Bradley, A. A., Schwartz, S. S., & Hashino, T. (2008). Sampling uncertainty and confidence intervals for the Brier score and Brier skill score. *Weather and Forecasting*, *23*(5), 992–1006. https://doi.org/10.1175/2007waf2007049.1

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics*, *14*(3), 1171–1179. https://doi.org/10.1214/aos/1176350057

Chernick, M. R., & LaBudde, R. G. (2011). *An introduction to bootstrap methods with applications to R* (p. 240).

Clark, M. P., & Slater, A. G. (2006). Probabilistic quantitative precipitation estimation in complex terrain. *Journal of Hydrometeorology*, *7*(1), 3–22. https://doi.org/10.1175/jhm474.1

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., et al. (2008). Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrologic models. *Water Resources Research*, *44*(12), W00B02. https://doi.org/10.1029/2007wr006735

Clark, M. P., & Vrugt, J. A. (2006). Unraveling uncertainties in hydrologic model calibration: Addressing the problem of compensatory parameters. *Geophysical Research Letters*, *33*(6), L06406. https://doi.org/10.1029/2005GL025604

Cornes, R. C., van der Schrier, G., van den Besselaar, E. J., & Jones, P. D. (2018). An ensemble version of the E-OBS temperature and precipitation data sets. *Journal of Geophysical Research: Atmospheres*, *123*(17), 9391–9409. https://doi.org/10.1029/2017jd028200

Coxon, G., Freer, J., Wagener, T., Odoni, N. A., & Clark, M. (2014). Diagnostic evaluation of multiple hypotheses of hydrologic behaviour in a limits-of-acceptability framework for 24 UK catchments. *Hydrological Processes*, *28*(25), 6135–6150. https://doi.org/10.1002/hyp.10096

Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R., & Smith, P. J. (2015). A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. *Water Resources Research*, *51*(7), 5531–5546. https://doi.org/10.1002/2014wr016532

Efron, B. (1992). Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *54*(1), 83–111. https://doi.org/10.1111/j.2517-6161.1992.tb01866.x

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, *37*(1), 36–48. https://doi.org/10.2307/2685844

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, *1*(1), 54–75. https://doi.org/10.1214/ss/1177013815

Everitt, B. S. (2002). *The cambridge dictionary of statistics* (2nd ed.). Cambridge University Press (ISBN: 0-521-81099-X).

Farmer, W. H., & Vogel, R. M. (2016). On the deterministic and stochastic use of hydrologic models. *Water Resources Research*, *52*, 5619–5633. https://doi.org/10.1002/2016WR019129

Fenicia, F., Savenije, H. H., Matgen, P., & Pfister, L. (2007). A comparison of alternative multiobjective calibration strategies for hydrologic modeling. *Water Resources Research*, *43*(3), W03434. https://doi.org/10.1029/2006wr005098

Fisher, R. A. (1920). Accuracy of observation, a mathematical examination of the methods of determining, by the mean error and by the mean square error. *Monthly Notices of the Royal Astronomical Society*, *80*, 758–770. https://doi.org/10.1093/mnras/80.8.758

Fowler, K., Peel, M., Western, A., & Zhang, L. (2018). Improved rainfall-runoff calibration for drying climate: Choice of objective function. *Water Resources Research*, *54*(5), 3392–3408. https://doi.org/10.1029/2017wr022466

Frei, C., & Isotta, F. A. (2019). Ensemble spatial precipitation analysis from rain gauge data: Methodology and application in the European Alps. *Journal of Geophysical Research: Atmospheres*, *124*(11), 5757–5778. https://doi.org/10.1029/2018jd030004

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrologic modeling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Gupta, H. V., Wagener, T., & Liu, Y. (2008). Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrological Processes*, *22*(18), 3802–3813. https://doi.org/10.1002/hyp.6989

Hampel, F., Ronchetti, E., Rousseeuw, P., & Stahel, W. (1986). *Robust Statistics, the approach based on influence functions*. Wiley.

Jolliffe, I. T. (2007). Uncertainty and inference for verification measures. *Weather and Forecasting*, *22*(3), 637–650. https://doi.org/10.1175/waf989.1

Kiang, J. E., Gazoorian, C., McMillan, H., Coxon, G., Le Coz, J., Westerberg, I. K., et al. (2018). A comparison of methods for streamflow uncertainty estimation. *Water Resources Research*, *54*(10), 7149–7176. https://doi.org/10.1029/2018wr022708

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*(3), W03S04. https://doi.org/10.1029/2005WR004362

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, *424*, 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., & Woods, R. A. (2020). A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments. *Water Resources Research*, *56*, e2019WR025975. https://doi.org/10.1029/2019WR025975

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 1217–1241. https://doi.org/10.1214/aos/1176347265

Lamontagne, L., Barber, C., & Vogel, R. M. (2020). Improved estimators of model performance efficiency for skewed hydrologic data. *Water Resources Research*, *56*, e2020WR027101. https://doi.org/10.1029/2020WR027101

Legates, D. R., & McCabe, G. J., Jr (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, *35*(1), 233–241. https://doi.org/10.1029/1998wr900018

Mansanarez, V., Renard, B., Coz, J. L., Lang, M., & Darienzo, M. (2019). Shift happens! adjusting stage-discharge rating curves to morphological changes at known times. *Water Resources Research*, *55*(4), 2876–2899. https://doi.org/10.1029/2018wr023389

Markstrom, S. L., Hay, L. E., & Clark, M. P. (2016). Towards simplification of hydrologic modeling: Identification of dominant processes. *Hydrology and Earth System Sciences*, *20*(11), 4655–4671. https://doi.org/10.5194/hess-20-4655-2016

Maurer, E. P., Wood, A. W., Adam, J. C., Lettenmaier, D. P., & Nijssen, B. (2002). A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States. *Journal of climate*, *15*(22), 3237–3251. https://doi.org/10.1175/1520-0442(2002)015<3237:ALTHBD>2.0.CO;2

McCuen, R. H., Knight, Z., & Cutter, A. G. (2006). Evaluation of the Nash–Sutcliffe efficiency index. *Journal of Hydrologic Engineering*, *11*(6), 597–602. https://doi.org/10.1061/(asce)1084-0699(2006)11:6(597)
Mizukami, N., Rakovec, O., Newman, A. J., Clark, M. P., Wood, A. W., Gupta, H. V., & Kumar, R. (2019). On the choice of calibration metrics for "high-flow" estimation using hydrologic models. *Hydrology and Earth System Sciences*, *23*(6), 2601–2614. https://doi.org/10.5194/hess-23-2601-2019
Mo, R., Ye, C., & Whitfield, P. H. (2014). Application potential of four nontraditional similarity metrics in hydrometeorology. *Journal of Hydrometeorology*, *15*(5), 1862–1880. https://doi.org/10.1175/jhm-d-13-0140.1
Montanari, A., & Koutsoyiannis, D. (2012). A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research*, *48*(9), W09555. https://doi.org/10.1029/2011wr011412
Moriasi, D. N., Gitau, M. W., Pai, N., & Daggupati, P. (2015). Hydrologic and water quality models: Performance measures and evaluation criteria. *Transactions of the ASABE*, *58*(6), 1763–1785. https://doi.org/10.13031/trans.58.10715
Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, *116*(12), 2417–2424. https://doi.org/10.1175/1520-0493(1988)116<2417:ssbotm>2.0.co;2
Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models. Part I—A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6
Nearing, G. S., Ruddell, B. L., Clark, M. P., Nijssen, B., & Peters-Lidard, C. (2018). Benchmarking and process diagnostics of land models. *Journal of Hydrometeorology*, *19*(11), 1835–1852. https://doi.org/10.1175/JHM-D-17-0209.1
Nelson, B. L., & Schmeiser, B. W. (1986). Decomposition of some well-known variance reduction techniques. *Journal of Statistical Computation and Simulation*, *23*(3), 183–209. https://doi.org/10.1080/00949658608810871
Newman, A. J., Clark, M. P., Craig, J., Nijssen, B., Wood, A., Gutmann, E., et al. (2015). Gridded ensemble precipitation and temperature estimates for the contiguous United States. *Journal of Hydrometeorology*, *16*(6), 2481–2500. https://doi.org/10.1175/jhm-d-15-0026.1
Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015
Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., & Nearing, G. (2017). Benchmarking of a physically based hydrologic model. *Journal of Hydrometeorology*, *18*(8), 2215–2225. https://doi.org/10.1175/JHM-D-16-0284.1
Papalexiou, S. M. (2018). Unified theory for stochastic modelling of hydroclimatic processes: Preserving marginal distributions, correlation structures, and intermittency. *Advances in Water Resources*, *115*, 234–252. https://doi.org/10.1016/j.advwatres.2018.02.013
Pool, S., Vis, M., & Seibert, J. (2018). Evaluating model performance: Towards a non-parametric variant of the Kling-Gupta efficiency. *Hydrological Sciences Journal*, *63*(13–14), 1941–1953. https://doi.org/10.1080/02626667.2018.1552002
Rakovec, O., Kumar, R., Mai, J., Cuntz, M., Thober, S., Zink, M., et al. (2016). Multiscale and multivariate evaluation of water fluxes and states over European river basins. *Journal of Hydrometeorology*, *17*(1), 287–307. https://doi.org/10.1175/jhm-d-15-0054.1
Reusser, D. E., Blume, T., Schaefli, B., & Zehe, E. (2009). Analysing the temporal dynamics of model performance for hydrologic models. *Hydrology and Earth System Sciences*, *13*(7), 999–1018. https://doi.org/10.5194/hess-13-999-2009
Ritter, A., & Muñoz-Carpena, R. (2013). Performance evaluation of hydrologic models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. *Journal of Hydrology*, *480*, 33–45. https://doi.org/10.1016/j.jhydrol.2012.12.004
Santos, L., Thirel, G., & Perrin, C. (2018). Pitfalls in using log-transformed flows within the KGE criterion. *Hydrology and Earth System Sciences*, *22*, 4583–4591. https://doi.org/10.5194/hess-22-4583-2018
Schaefli, B., & Gupta, H. V. (2007). Do Nash values have value? *Hydrological Processes*, *21*, 2075–2080. https://doi.org/10.1002/hyp.6825
Seibert, J. (2001). On the need for benchmarks in hydrological modelling. *Hydrological Processes*, *15*(6), 1063–1064. https://doi.org/10.1002/hyp.446
Seibert, J., Vis, M. J., Lewis, E., & Meerveld, H. V. (2018). Upper and lower benchmarks in hydrological modelling. *Hydrological Processes*, *32*, 1120–1125. https://doi.org/10.1002/hyp.11476
Tang, G., Clark, M. P., & Papalexiou, S. M. (2021). SC-Earth: A station-based serially complete earth dataset from 1950 to 2019. *Journal of Climate*, *34*, 1–47. https://doi.org/10.1175/jcli-d-21-0067.1
Tang, G., Clark, M. P., & Papalexiou, S. M. (2021). The use of serially complete station data to improve the temporal continuity of gridded precipitation and temperature estimates. *Journal of Hydrometeorology*, *22*(6), 1553–1568. https://doi.org/10.1175/jhm-d-20-0313.1
Tang, G., Clark, M. P., Papalexiou, S. M., Newman, A. J., Wood, A. W., Brunet, D., & Whitfield, P. H. (2021). EMDNA: Ensemble meteorological dataset for North America. *Earth System Science Data*, *13*, 3337–3362. https://doi.org/10.5194/essd-13-3337-2021
Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, *43*(1), W01413. https://doi.org/10.1029/2005wr004723
Van Werkhoven, K., Wagener, T., Reed, P., & Tang, Y. (2009). Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models. *Advances in Water Resources*, *32*(8), 1154–1169. https://doi.org/10.1016/j.advwatres.2009.03.002
Vogel, R. M. (2017). Stochastic watershed models for hydrologic risk management. *Water Security*, *1*, 28–35. https://doi.org/10.1016/j.wasec.2017.06.001
Vogel, R. M., & Shallcross, A. L. (1996). The moving blocks bootstrap versus parametric time series models. *Water Resources Research*, *32*(6), 1875–1882. https://doi.org/10.1029/96wr00928
Wang, Z., & Bovik, A. C. (2009). Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine*, *26*(1), 98–117. https://doi.org/10.1109/msp.2008.930649
Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (Vol. 100). Academic press.
Willmott, C. J. (1981). On the validation of models. *Physical Geography*, *2*(2), 184–194. https://doi.org/10.1080/02723646.1981.10642213
Wright, D. P., Thyer, M., Westra, S., Renard, B., & McInerney, D. (2019). A generalised approach for identifying influential data in hydrological modelling. *Environmental Modelling & Software*, *111*, 231–247. https://doi.org/10.1016/j.envsoft.2018.03.004
Yapo, P. O., Gupta, H. V., & Sorooshian, S. (1998). Multi-objective global optimization for hydrologic models. *Journal of Hydrology*, *204*(1–4), 83–97. https://doi.org/10.1016/s0022-1694(97)00107-8
Ye, L., Gu, X., Wang, D., & Vogel, R. M. (2021). An unbiased estimator of coefficient of variation of streamflow. *Journal of Hydrology*, *594*, 125954. https://doi.org/10.1016/j.jhydrol.2021.125954