Details of revision based on comments from referees

This documents how we addressed the comments in the revised manuscript. The original review comments are in black, and our responses in red. We also made additional textual revision throughout the text.

RC1

The paper compares the KGE, NSE and a peak flow signature as objective function for the calibration of 2 hydrological models. The paper is well written and clear. However, it does not lead to new results, and the suggestion to abandon NSE in favour of KGE is not well motivated. These points are further elaborated below.

The authors are strongly in favour of KGE vs NSE, as clearly appears from statements such as "*Squared error metrics, such as Nash Sutcliffe Efficiency (NSE) and Mean Square Error (MSE), have historically been thought to be useful to reduce simulation errors associated with high flow values (Oudin et al., 2006; Price et al., 2012; Seiller et al., 2017; de Boer-Euser et al., 2017). 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*".

One cannot expect NSE to have properties that it is not designed to have, and it would be fair to use such metric in a way that is meaningful and that reflects the theory behind it.

The use of sum of squared errors and its rescaled variants is common in statistics, and can be related to precise assumptions about the error. In particular, such objective functions follow the assumption of Normal, zero mean, iid residuals. This is among the simplest assumptions one can make, although often inappropriate, as widely discussed. The properties of a model calibrated using NSE should be considered within the context of this theory. The fact that a deterministic model calibrated using NSE will underestimate the variability of the flow is NOT a design flaw of the NSE. It is a characteristic that follows from theory. From theory one can also easily see that it comes to no surprise that the statistics of the deterministic model don't match the statistics of the observed data. They will not match by design. In particular, if the assumption behind squared error metrics is that Qobs=Qmod+eps (with eps N(0,sigma)), it is obvious that the statistics of Qmod are different from the statistics of Qobs. The statistics of Qobs should be compared to the statistics of Qmod+eps. This explains also why, for example, var(Qobs)>var(Qmod). Of course it is, since var(Qobs)=var(Qmod +err)= var(Qmod)+var(err). I can see that the approach of correctly comparing modelled and observed statistics (ie accounting for the error) is almost never followed in the community. This has led to the wrong perception that NS and related metrics somehow don't work.

Therefore, before recommending to switch to other metrics, I would propose the 'old' metrics be tested fairly. Based on this, I have the following suggestions for this paper: Don't provide poorly grounded indications such as that "squared error type metrics are not suitable for model calibration when the application requires robust high flow performance". NSE and KGE are based on different assumptions, and they should be compared fairly. Even if the KGE results into better performance, one should still note that NSE can be related to properties of the errors, which can be tested and changed if necessary (e.g. one can use the NSE of the sqrt of the flow to reduce heteroscedasticity).

At present I don't see the novelty of this paper. Most of the statements about the perceived qualities of KGE (part of them debatable, as I explained), are already given in other papers. Conclusion 1 is expected by design of the calibration metrics. Conclusion 2 is unclear. Conclusion 3 is highly debatable as explained.

References:

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

Kavetski, D., F. Fenicia, P. Reichert, and C. Albert (2018), Signature-Domain Calibration of Hydrological Models Using Approximate Bayesian Computation: Theory and Comparison to Existing Applications, Water Resour Res, 54(6), 4059-4083.

We very much appreciate comments coming from different perspective related to modeled flow metrics. We completely agree with the main comment – that sum-of-squared error metrics commonly used in optimization reduce the variance by design, and representing the flow statistical moments and extremes requires stochastically simulating the error term. We carefully reviewed the paper by Famer and Vogel (2016) who discuss about the stochastic estimation of model errors and effect of residuals on high flow metrics. The ideas illustrated in the Farmer and Vogel (2016) paper help us better frame our contribution.

In the revised discussion section (section 5.2 P8-9), we analyze the distributions of errors for models calibrated using three objective functions (KGE/NSE/APFB), as well as examining the flow time series metrics (mean and variability) and high flow metrics for deterministic KGE/NSE/APFB calibrated flow and ensemble flows based on stochastically generated error added to respective calibrated streamflow simulations. As you see in the discussion, we observe that the ensemble of residual reintroduced flows does improve the metrics (mean, variances, and high flow statistics) regardless of residual distribution of the deterministic flow. However, the dynamical property (i.e., temporal pattern) deteriorates due to lack of temporal correlation in the synthetic flow sequences. The method for stochastic residual generation uses random sampling based on flow magnitudes and requires incorporation of auto correlation properties.

While we can arrive at similar conclusions to Farmer and Vogel (2016), we feel that obtaining improved deterministic flow simulation through model calibration is important because of improvement of application specific flow metrics through improving the magnitude, variability, temporal correlation. And we feel that impacts of performance metric choice on deterministic flow metrics are still not well appreciated by the broader community. We hope that our paper provides additional explanations of unintended consequences of model calibration decisions.

We have slightly revised the conclusions in the revised manuscript. Our main point however still remains – alternatives to sum-of-squared error metrics can improve the deterministic component of the model simulations, especially for high flows. This is important since most hydrologic modeling applications only consider the deterministic component.

RC2

1 OVERALL RECOMMENDATION

The manuscript addresses the important topic of the choice of calibration metrics (CM) to be used for rainfall-runoff modeling, and presents results obtained on 492 US catchments. I found the paper interesting, including relevant references. If the presented results are not highly original, the paper is, in his present form, an excellent illustration of the limitation of the use of Nash and Sutcliffe efficiency metric (NSE, 1970) for model calibration. Nevertheless, I do have major comments on the used dataset, the applied methodology and the discussion part. Thus, I recommend to accept the manuscript in HESS with major revisions detailed below.

2. GENERAL COMMENTS

2.1 Description of the studied catchments
Even though the objective of such "large-sample hydrology paper" is not to present results obtained on a limited number of catchments having the same hydrometeorological characteristics but to have general conclusions on rainfall-runoff modeling, I think the diversity of the studied catchments has to be addressed and quantified.

This description is lacking right now in the paper. A presentation of the general characteristics of the studied catchments should be added in the paper, in order to understand the variability of catchments characteristics (catchment area, runoff coefficient, mean annual solid precipitation, etc.), especially in the context of flood modeling: what are different flood processes and dynamics included within this catchments sets (flash floods, snowmelt floods, rain-on-snow floods, groundwater floods, etc.)? Moreover, the timestep considered in the two rainfall-runoff models is not stated in the paper and should be mentioned. Are the models working at daily timestep? Is this timestep consistent with the flood dynamics of every studied catchment?

Yes, one of main objectives of large sample basin study is to generalize the conclusions drawn regarding hydrologic modeling evaluations (Gupta et al 2014). This manuscript used a subset (492 out of 671) the catchments presented by Addor et al., (2018), who describe in detail the variability of climate/geophysical/hydrologic characteristics for the 671 US catchments. Our basin selections are also spread over the CONUS; therefore, distributions of basin characteristics are similar to Addor et al., (2018). We decided to avoid repetitive summaries and figures.

We looked at spatial pattern of the model skills in addition to the distributions (See Figs R1 and R2). There is little distinct spatial pattern in the APFB (%bias Qpeak in Fig R2). This indicates that catchment characteristics have less effects than the performance metrics P6, L25. We mentioned this text in L-235-236 (not shown in Figures in the revised manuscript).

As for the time step, we performed daily simulation for calibration in this study as we stated in P5, L10. Though calibrated parameter values may not be consistent for different temporal resolutions, the trend in calibration performances across the different performance metrics should be preserved regardless of the time steps. Moreover, the theory of algebraic decomposition of NSE is independent of the time step.

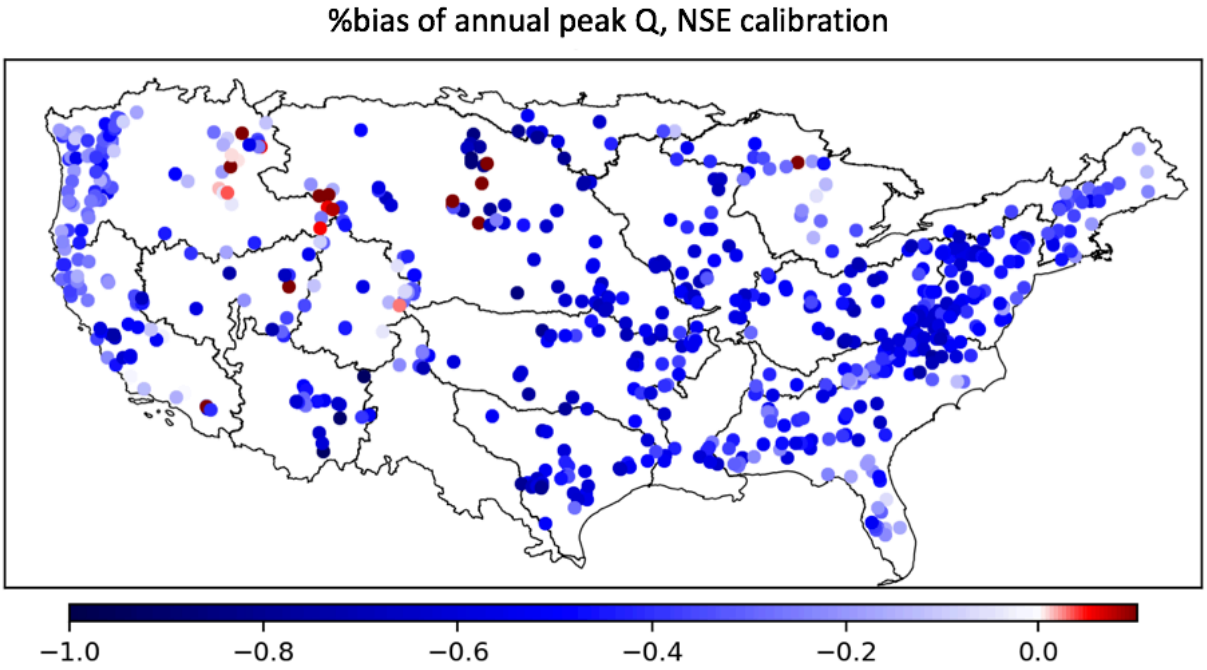


Figure R1. %bias of annual peak flow (APFB) distribution from VIC model calibrated with NSE

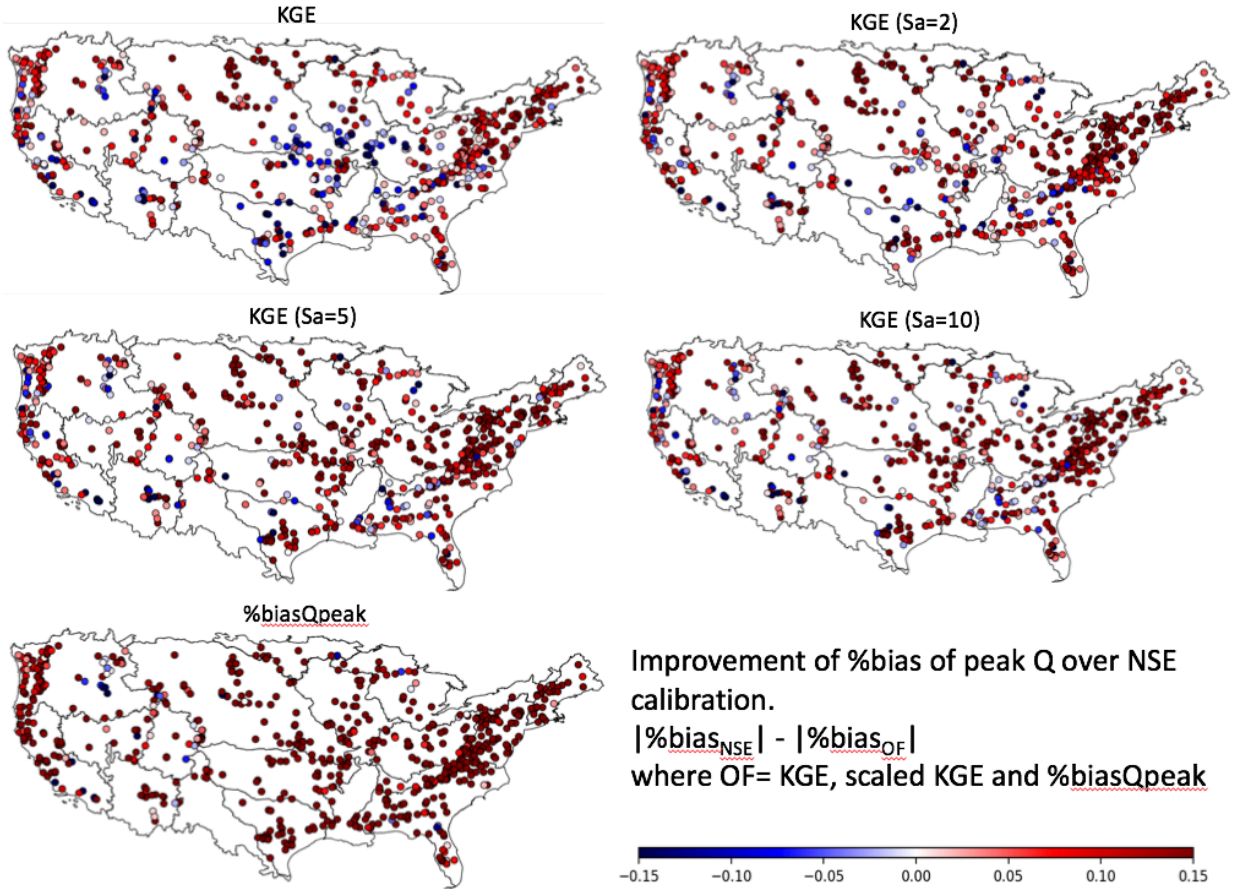


Figure R2. Improvement of %bias of annual peak flow (APFB) over NSE calibration for KGE and APFB calibrations. The model is VIC. Red indicate the KGEs or APFB improve over NSE APFB while blue indicate the opposite.

2.2 Split-sample test

For every catchment, the calibration and validation periods are the same time-periods, 1999-2008 and 1989-1999, respectively. I think that performing a basic split and sample test (Klemeš, 1986) on each catchment would be particularly interesting in this context, especially to address temporal (in)stability of parameter sets obtained with particular CM (topic partially addressed page 6, line 12).

It may not be clear enough in the text but we did use a split-sample method i.e., calibration period during 1999-2008 and evaluation during 1989-1999 (P4, Line 29-31 in original manuscript). The exception is the evaluation of annual peak flow error where we combined the two periods (P6, Line 24 in original manuscript) to have increased sample size. Please see the slightly revised version on P5, L12-14.

2.3 List of the studied CM

The paragraph listing the studied CM (page 5, lines 5 to 16) is unclear and would be easier to understand if a list (or table) of the five studied CM was added.

We re-wrote the paragraph listing the metrics used for calibration and some descriptions in P 5, L 21–30.

2.4 “Application-specific” or “hydrologic signature”?

From page 5 to the end of the paper, APFB is named as an “application-specific” metric, while being introduced as an “hydrologic signature” (see definition of “hydrologic signature” in the paper introduction, page 2 lines 10 to 24) in the paper objective presentation. What is the difference between an “application specific” and a “hydrologic signature” CM in this context? Finally, is APFB an “application specific” or an “hydrologic signature” CM? Could you address this point?

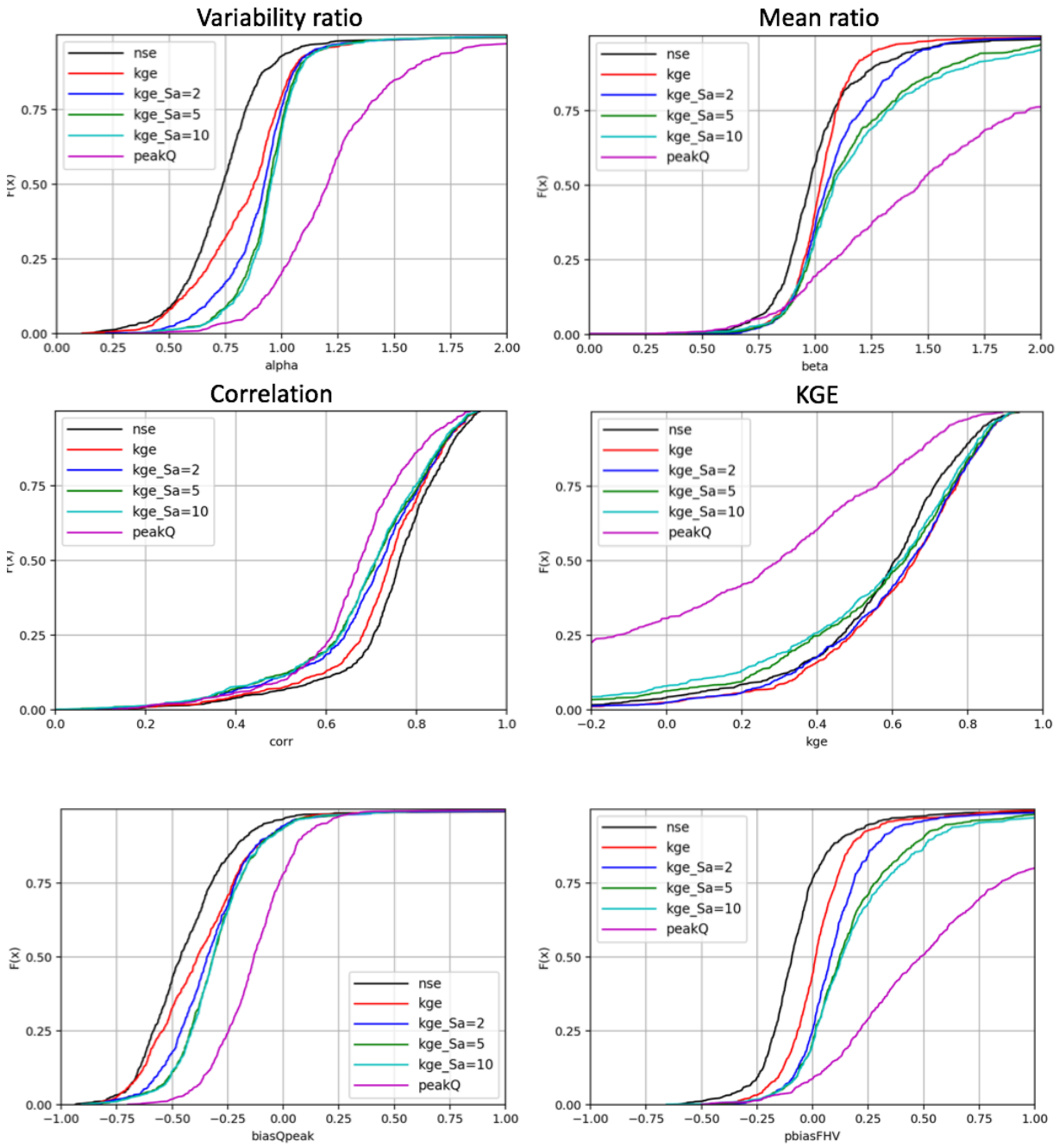
We clarified the difference in the definition between “hydrologic signature” and “application-specific metric”. The “hydrologic signature” is a metric derived from hydrologic variable time series to imply how a specific hydrologic process behaves, as expressed by the model simulation (or observation) for better understanding of the processes. So, it is process-oriented metric. Please see P2, L13-19. On the other hand, while the “application-specific metric” can be a hydrologic signature, it is a metric used by water managers for specific decision-making. Therefore, it can be simple bias of monthly or annual time series, which can be used for long term hydroclimate assessment purpose (not worry about hourly, or daily variability). Please see P2, L29-33. The hydrologic signature can be used as an application specific metric (please see P2, L33-35.

2.5 Impact of the KGE scaling factors

The limited impact of the different KGE scaling factors used in the paper is very little discussed, while being particularly interesting. This point has to be discussed in the paper. Moreover, why not trying another combination with a larger variability ratio scaling factor, such as (Sr=1, Saplha=20, Sbeta=1), to assess a potential significant improvement of annual peak flow bias? What about another test with (Sr=0.1, Saplha=1, Sbeta=0.1) or even (Sr=0, Saplha=1, Sbeta=0)?

We actually tested a few extreme cases: (Sr=1, Saplha=10, Sbeta=1) and (Sr=0, Saplha=1, Sbeta=0). It did not show improvement of annual peak flow. These results (all the statistics) did not change significantly

(statistically at 95%) from the case (Sr=1, Saplha=5, Sbeta=1) as shown below. No textual change has been made for this concern.



2.6 Figures

In general, the presentation of the figures could be improved for a better understanding:

- The performance metrics plotted have different names in the axis labels and in the figure legends (e.g. it is not explicit that "%biasFHV" is equal to "percentage bias of flow volume above 80 percentile flow duration curve" in the Figure 4) ;

- The name/typography of several performance metrics is changing over figures (“%bias Qpeak” on Figure 2 but “%biasQpeak” on Figure 4) ; - Why not using Greek letters in Figure 2 x-axis?

- The link between the five CM and the figure legend is never clearly stated, and for example, the reader has to guess that “kge_2alpha” is equal to (Sr=1, Saplha=2, Sbeta=1).

- How boxplots have been constructed? What are the outlier points plotted below and over the boxplots?

We corrected the figures based on the above comments. We use APFB to denote percent bias in annual peak flow (%bias Qpeak)

3 SPECIFIC COMMENTS
1. Page 4, line 27: please change (Maurel et al., 2002) into Maurel et al. (2002).

We corrected the above error in L-176.

2. Page 13, figure 2: please state in the figure legend that results presented in this figure are obtained with the VIC model.

We corrected the above issue in Figure 2 caption.

4 REFERENCES
Klemeš, V., 1986. Operational testing of hydrological simulation models. Hydrological Sciences Journal 31, 13. https://doi.org/10.1080/02626668609491024.

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

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

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

[1] National Center for Atmospheric Research, Boulder CO, USA
[2] UFZ-Helmholtz Centre for Environmental Research, Leipzig, Germany
[3] Czech University of Life Sciences, Prague, Czech Republic
[4] Department of Hydrology and Atmospheric Sciences, The University of Arizona, Tucson, Arizona, USA

Corresponding author: N. Mizukami (mizukami@ucar.edu)

Abstract

Calibration is an essential step for improving the accuracy of simulations generated using hydrologic models, and a key modeler decision is the selection of the performance metric to be optimized. It has been common to used 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 estimation of high flows. However, we find 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 daily step, the "Variable Infiltration Capacity" model (VIC) 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 "hydrologic signature" metric that focuses on annual peak flows. As expected, the application specific 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 result in underestimation of observed flow variability. MeanwhileOn 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. EnsembleEnsembles of the Sstochastically generated ensemble simulations stochastically generated based on remaining residuals canshow ability to recoverimprove some of the metrics regardless of the deterministictheir performances of theduring the model calibration processs. However, it should beis emphasized that gettingobtaining the correct fidelity of streamflow dynamics of from the deterministically calibrated models areis still important andforas it may improvee the high flow metrics (for the right reasons). Overall this work highlights the need for a fuller deeper understanding of performance metric behavior and design in relation to the desired goals of model calibration.

1 Introduction

[1] Computer-based hydrologic, land-surface, and water balance models are used extensively to generate continuous long-term hydrologic simulations in support of water resources 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 system response (e.g., streamflow).

[2] A key decision in model calibration is the choice of performance metric (also known as "objective function") that measures the goodness of fit between the model simulation and system observations, because 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; Price et al. 2012; Pushpalatha et al. 2012; Wöhling et al. 2013; 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.

[3] 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. 2017*) has been gaining popularity for hydrologic model calibration *(Westerberg et al. 2011; Yadav et al. 2007; Shafii and Tolson 2015; Kavetski et al. 2018*). A hydrologic signature is a ~~statistic~~ 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, flow correlation structure, etc.), in such a way that it is informative regarding a specific hydrologic process of a catchment (*Yilmaz et al. 2008*).

[4] The use of hydrologic signatures to form metrics for model calibration requires selecting 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.

[5] In general, water resources 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*). ~~Therefore, the hydrologic model is often optimized for a specific target metric that is the most relevant to the application.~~ Hereafter ~~we call~~ this metric is referred to as 'application specific metric'. ~~Note~~It is worth noting that the application specific metric can be a hydrologic signature metric. For example, ~~volume of~~ high flow volume based on the flow duration curve~~which volume based on flow duration curve~~

characterizecharacterizes the surface flow processes and and may be interest for estimation of flood frequency.

[6] 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:

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

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

[7] 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; *Samaniego et al. 2010; Kumar et al. 2013)* 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), 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 using based on metric -(1) NSE, (2) Kling-Gupta Efficiency (KGE) and its variants, and (3) Annual Peak Flow Bias (APFB) – with focus on understanding and evaluating the appropriateness of different metrics to capture observed high flow behaviors across a diverse range of U.S. basins. We also perform and discuss the implications of the choice of different calibration metrics based on stochastic ensemble simulations generated based on remaining model residuals.

[8][7]

[9][8] 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 for 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 Section 5. Concluding remarks are provided in Section 6.

2 Motivation

[10][9] The earliest development of a metric used for model development is by (Nash and Sutcliffe (1970), who proposed assessing Mean Square Error (MSE) relative to the observation mean; Nash Sutcliffe Efficiency (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 squared error metrics, such as Nash Sutcliffe Efficiency (NSE) and Mean Square Error (MSE), have been predominantly used for model evaluation as well as for model calibration. historically 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(*Oudin et al. 2006; Price et al. 2012; Seiller et al. 2017; de Boer-Euser et al. 2017*). 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 (*Oudin et al. 2006; Price et al. 2012; Seiller et al. 2017; de Boer-Euser et al. 2017*). Via an algebraic reformulation of 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 ~~only in~~is the ideal but unachievable ~~situation~~ one when the correlation is 1.0. They further show that ~~this results~~the consequence is ~~in~~ a tendency to underestimate high flows while overestimating low flows (see Fig 3 in *Gupta et al 2009*).

~~[11]~~[10] 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 with 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). Figures 1(b-d) show clearly 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 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.

~~[12]~~[11] 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 Kling-Gupta Efficiency (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:

$$KGE = 1 - \sqrt{[S_r(r-1)]^2 + [S_\alpha(\alpha-1)]^2 + [S_\beta(\beta-1)]^2} \qquad \text{Eq. (1)}$$

$$\alpha = {}^{\sigma_s}/_{\sigma_o}, \quad \beta = {}^{\mu_s}/_{\mu_o}$$

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.

~~[13]~~[12] 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 %bias in annual peak flows.

3 Models, Datasets and Methods

~~[14]~~[13] We use~~d~~ two hydrologic models; VIC and mHM. 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 large-sample predictability benchmark study~~ies~~ (*Newman et al. 2017*). The mHM model has been shown to provide robust hydrologic simulations over both Europe and the US (*Kumar et al. 2013a; Rakovec et al. 2016a*) and is currently being used in application studies (*Samaniego et al. 2018*). We use daily observed streamflow data at the HCDN basins and

daily basin meteorological data from *Maurer et al. (2002)* for the period 1980 through 2008, as compiled by the CONUS large sample basin dataset over a wide range of climate regimes (*Addor et al. 2017; Newman et al. 2014*). Interested readers may refer to *Addor et al. 2017 and Newman et al. 2014* for more details and insights into different physiographic and hydro-climatic chcaracteristics of the study basins. The use of thise large sample dataset helps to obtain moreis recommended to obtain general and statistically robust conclusions (*Gupta et al. 2014*). Both models are run at a daily time scalestep, and each of themmodel is calibrated separately Each of the models was independently calibrated at daily step to each of the for each of the 492 HCDN study basins (see Fig. 1a for the basin locations) across the CONUS domain using several different performance metrics. We use observed streamflow data at the HCDN basins for the period 1980 through 2008, and daily basin meteorological data from *Maurer et al. (2002)*, as compiled by the large sample basin dataset (*Addor et al. 2017; Newman et al. 2014*). The use of this large sample dataset helps to obtain more general and statistically robust conclusions (*Gupta et al. 2014*). We split use a split-sample approach (*Klemes 1986*) for the model evaluation. Tthe hydrometeorological data is split into a calibration period (October 1, 1999 - September 30, 2008) and an evaluation period (October 1, 1989 - September 30, 1999), and usedwith a prior 10-year warm-up when computing the statistics for each period.

[15][14] The model parameters calibrated for each model are the same as previously discussed: VIC (*Mizukami et al. 2017; Newman 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 waisis identically configured for each of the 492 basins, and b. Both models used 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.

[16][15] Optimization was is performed using the Dynamically Dimension Search (DDS; *Tolson and Shoemaker 2007*) algorithm. Five performance metrics were are used for the calibration/evaluation purpose: 1) KGE, 2) KGE-2α, 3) KGE-5α, and 4) APFB (Annual Peak flow bias) and 5) NSE. The first tthree metrics arebased on 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 (Fig. 1c), we explore the impact of rescaling the variability error term in Eq. (1).with varying scaling factors to emphasize different components, one being an application-specific high flow metric, The forth metric, APFB, isFor our application-specific high flow metric, we use the Annual Peak Flow Bias (*APFB*) measure defined as:

$$\mathrm{APFB} = \sqrt{\left[\mu_{peakQ_s}/\mu_{peakQ_o} - 1\right]^2} \qquad (2)$$

where μ_{peakQ_s} is the mean of the simulated annual peak flow series and μ_{peakQ_o} is the mean of the observed annual peak flow series. and ourFinally, we took NSE is regard as a benchmark performance metric, and compared and contrasted the simulations based on other performance metrics being the NSE. For KGE, historically, the most common choice of scaling factor for hydrologic model calibration has been to set all of them to unity and, to the best of our knowledge, scaled KGE variants (i.e., with non-unity scaling factors) have not been well studied.

Historically, tThe 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 best and, to the best of our knowledge , scaled KGE variants (i.e., with non-unity scaling factors) have not been well studied so-far. 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), by using three formulations of KGE with (S_r, S_α and S_β) = (1,1,1), (1,2,1), and (1,5,1). 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

~~[17]~~[16] First, we focus on the general overall performance for the daily streamflow ~~performance~~ simulations as measured by the performance metrics used. Figures 2 and 3 show the cumulative distributions of ~~evaluation period~~ 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), (c) r (linear correlation) for VIC (Fig. 2) and mHM (Fig. 3). Considering first the result obtained using KGE, ~~. For~~ both models, at the median values of the distributions, ~~use of KGE improves~~show improvement in variability error score by approximately 20% over that obtained using the NSE-based calibration score (Figs. 2a and 3a)~~; however, . T~~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 mHM. Interestingly, the VIC results are more sensitive than 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).

~~[18]~~[17] ~~For both models, t~~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 (α=1) results in 3-10% lower NSE than those obtained ~~when calibrating~~ with the NSE-based calibration case (plots not shown). This is consistent with the expectation that an~~, because~~ improvement in the variability error score (α closer to unity) ~~is known to cause a reduction in~~leads to deterioration in the NSE ~~optimality~~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

~~[19]~~[18] 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 (Figure 4 a and b), while use of NSE produces the worst peak flow estimates. Simply switching from NSE to KGE improves the ~~percentage bias of peak flow~~APFB by approximately 5% for VIC and 10% for mHM at the median value during evaluation. Improvement of APFB occur at over 85-% of 492 basins ~~across CONUS~~ 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 ~~the bias of annual peak flow~~APFB is used as the objective function (see the results from mHM; Figure 4 a 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.

[19] Figure 4 c and d show the high flow simulation performance in terms of another high flow related metric – the percent ~~%~~bias ~~of~~ in the runoff volume ~~of~~ 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 ~~property~~ 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~~model-based results also exhibit larger variability in the FHV bias across the study basins.

~~[20]~~

4.3 Implication for flood frequency estimation

~~[21]~~[20] Annual peak flow estimates are generally used directly in the~~for~~ flood frequency analysis. Figure 5 shows estimated daily flood magnitudes at three return periods (5-, 10-, 20-yr) 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 20-yr (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 validation periods.

~~[22]~~[21] Figure 5 shows results that are consistent with Figure 4, although more outlier basins ~~were~~ 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 ~~the~~ three return periods for both models. In ~~Particular~~particular~~ly~~, 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 ~~the %bias of annual peak flow~~APFB is less than 5% at the median value for mHM calibrated with APFB (Figure 4), the 20-yr flood magnitude is underestimated by almost 20% at the median (Figure 5). Also, the degree of underestimation of flood magnitude becomes ~~greater~~ larger with longer return periods.

5 Discussion

~~[23]~~ ~~Overall, while~~While ~~-~~both models show fairly similar trends in skill for each performance metric ~~used for calibration~~, it is clear from our large sample study of 492 basins that the absolute performance of VIC is ~~always poorer~~~~rather~~ inferior ~~than~~to that of mHM, irrespective of choice of evaluation metric. A full investigation of~~f~~n to why VIC does not perform ~~as well as~~at the same level of mHM is clearly of interest, but ~~-this~~ is left for ~~the~~ 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. *Mendoza et al. 2015*~~;~~ ~~*Cuntz et al. 2016*~~)~~. ~~In~~ Below, ~~:~~

[22] ~~In this section~~ we discuss~~-~~ our results in the context of ~~of the~~ usage of different performance metrics, ~~on~~in regard to remaining aspects of model errors, and ~~suggest~~provide suggestions for potential improvement of the high flow ~~metric~~simulations related metrics~~s~~.

5.1 Consideration of application specific metric

~~[24]~~ 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~~, especially for VIC~~. 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-yr return period (top panels) and 5-yr return period

(bottom panels) are related to variability error and bias of annual peak flow estimates. As expected, the more extreme (20-yr return period) flood estimates are more strongly correlated with estimates of the variability of annual peak flows than with the 20-yr bias of the annual peak flow series. For the less extreme (5-yr return period) events, this trend is flipped and flood magnitude errors are more correlated with the bias.

[23] ~~Overall, while both models show fairly similar trends in skill for each performance metric used for calibration, it is clear from our large sample study of 492 basins that the absolute performance of VIC is always poorer than that of mHM, irrespective of choice of evaluation metric. A full investigation of why VIC does not perform as well as 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 include hard-coded parameters in other more complex models (e.g. *Mendoza et al. 2015; Cuntz et al. 2016*).~~

5.2 Consideration of model residuals

[24] 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, ~~always~~residuals always remain after the model calibration because the ~~calibrated~~ 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 and so on). ~~Here~~In the context of this study for the high flow simulations~~–~~, lets focus~~ing~~ on the flow variability (i.e., variance) component for observation and model simulations, which can expressed by ~~which is related to high flow simulation, the~~ following equation ~~is considered~~:

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

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

[25] To gain more insight~~s~~ into this topic, ~~Here~~ we examine how ~~the~~stochastic~~ally treatment of the~~generated residuals,~~once~~ once re-introduced to the simulated flows, can affect the performance ~~flow m~~metrics. We consider three performance metrics for this analysis: NSE, KGE, and APFB. ~~First,~~ 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 Figure 2d and 3d) ~~-~~produces wider residual distributions. Following the method ~~by~~of ~~(~~*Farmer and Vogel (2016) and~~;~~ Bourgin et al. (2015)*, 100 sets of synthetic residual time series (ε) ~~for the validation period~~ 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 ~~formed~~ based on the flow magnitude. For each of the 100 residual ~~reintroduced~~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. Similar to ~~(~~*Farmer and Vogel (2016)*~~results~~, high flow volume error (percent bias of FHV) and APFB

computed with residual incorporated flow series also improve ~~over~~ compared to the deterministic flow series from the calibrated models (Figure 9).

[25][26] ~~In summary, I~~~~it seems that the~~The quality of the original deterministic flow simulated by the hydrologic models has little effect on the ~~flow~~performance metrics based on the ensemble of residual ~~added~~augmented flows ~~(Figure ??)~~. Since the stochastically generated ensembles do not ~~count~~account for temporal correlation, every ensemble has reduced correlation and deteriorated ~~time series error metrics (e.g., NSE and KGE metrics)~~. However, the error metrics related to the flow duration curve (APFB) is not affected by the lack of correlation because metrics ~~using~~based on FDC do not ~~account for~~preserve information regarding the temporal sequence. ~~T~~Although residual ~~reintroduced~~augmented flow ~~ensemble~~time-series enhances some of flow metrics, the (temporal) dynamical~~l~~ pattern is not reproduced. These observations point toward the need for careful investigation ~~on~~in interpreting the improvement in model skill ~~–~~ especially when different error metrics are taken into account. ~~therefore, modeler should be aware of how model skills are improved.~~

6 Conclusions

[26][27] The use of large sample catchment calibrations of two different hydrologic models with ~~five~~ several performance metrics enables us to make robust inferences regarding the effects of the calibration metric on the ability to infer ~~extreme (~~high flow~~)~~ events. Here, we have focused on improvement in representation of annual peak flow ~~estimates~~estimates, as they are important for flood frequency magnitude estimation. ~~Our calibration study supports the notion of *Gupta et al. (2009)* that squared error type metrics are not suitable for model calibration when the application requires robust high flow performance.~~ We draw the following conclusions from the analysis presented in this paper:

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

~~1.~~2. Calibration with KGE improves performance as assessed by high flow metrics through improvement of time~~-series~~dependent metrics (~~mean,~~e.g., ~~–~~variability error score). Adju~~sting~~stment of the ~~-~~scaling factors ~~on~~related to the different ~~the~~ KGE components (bia~~ss,~~, variability, and ~~error,~~ correlation terms) can further assist the model simulations in matching certain aspects of flow characteristics. The degree of improvement is, however, ~~is~~ model dependent ~~, though sensitivity depends on the models~~.

~~2.~~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.

~~3. The ability to adjust weighting on bias, variability, and correlation makes KGE a versatile performance metric that can be used to improve model-based estimation of high flow related hydrologic signatures.~~

[27][28] Given that *Gupta et al. (2009)* shows 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~~, that are more complex,~~ we can expect that other models would exhibit similar results when using KGE or ~~a~~ its scaled variant. ~~It~~ If choosing to use an application specific metrics, it seems clear that careful thought needs to be given to the design of ~~application specific~~the metric~~s~~ if we are to obtain good performance for both the target metric (used for calibration) and other related metrics (used for evaluation). This is ~~more), so~~important since we wish ~~as~~ to increase confidence in the robustness and transferability of the calibrated model – an ~~the~~ issue that ~~. This issue~~ needs to be examined in more detail.

7 Code Availability

~~[28]~~[29] Model calibration was performed using MPR-flex available at https://github.com/NCAR/mpr-flex/tree/direct_calib for VIC. mHM is calibrated with the MPR strategy implemented in the mHM. Hydrometeorological data are obtained from a part of Catchment Attributes and Meteorology for Large-sample Studies (CAMELS; *Newman et al. 2014; Addor et al. 2017*). Analysis and plotting codes are available at https://github.com/nmizukami/calib4ffa~~https://github.com/nmizukami/calib4ffa/blob/master/ffa.ipynb~~.

Acknowledgements

We thank two anonymous referees for their constructive comments and Dr. Ding for his short comment on NSE. The comments help 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. This work was financially supported by the U.S Army Corps of Engineers Climate Preparedness and Resilience program.

References

Addor, N., A. J. Newman, N. Mizukami, and M. P. Clark, 2017: The CAMELS data set: catchment attributes and meteorology for large-sample studies. *Hydrol. Earth Syst. Sci.*, **21**, 5293–5313, doi:10.5194/hess-2017-169. https://www.hydrol-earth-syst-sci.net/21/5293/2017/hess-21-5293-2017.html.

Bergström, S., 1995: The HBV model. *Compute Models of Watershed Hydrology*, V.P. Singh, Ed., Water Resouces Publications, Highlands Ranch Co.

de Boer-Euser, T., and Coauthors, 2017: Looking beyond general metrics for model comparison -- lessons from an international model intercomparison study. *Hydrol. Earth Syst. Sci.*, **21**, 423–440, doi:10.5194/hess-21-423-2017. https://www.hydrol-earth-syst-sci.net/21/423/2017/.

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

Elsner, M., and Coauthors, 2010: Implications of 21st century climate change for the hydrology of Washington State. *Clim. Change*, **102**, 225–260, doi:10.1007/s10584-010-9855-0. http://dx.doi.org/10.1007/s10584-010-9855-0.

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

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

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

Gupta, H. V., S. Sorooshian, and P. O. Yapo, 1998: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resour. Res.*, **34**, 751–763, doi:10.1029/97wr03495. http://dx.doi.org/10.1029/97WR03495.

Gupta, H. V., K. J. Beven, and T. Wagener, 2006: Model Calibration and Uncertainty Estimation. *Encycl. Hydrol. Sci.*, doi:doi:10.1002/0470848944.hsa138. https://doi.org/10.1002/0470848944.hsa138.

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

——, H. Kling, K. K. Yilmaz, and G. F. Martinez, 2009: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.*, **377**, 80–91, doi:http://dx.doi.org/10.1016/j.jhydrol.2009.08.003. http://www.sciencedirect.com/science/article/pii/S0022169409004843.

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

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

https://doi.org/10.1002/2017WR020528.

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

Kumar, R., L. Samaniego, and S. Attinger, 2010: The effects of spatial discretization and model parameterization on the prediction of extreme runoff characteristics. *J. Hydrol.*, **392**, 54–69, doi:http://dx.doi.org/10.1016/j.jhydrol.2010.07.047. http://www.sciencedirect.com/science/article/pii/S0022169410004865.

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

Kumar, R., L. Samaniego, and S. Attinger, 2013b: Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. *Water Resour. Res.*, **49**, 360–379, doi:10.1029/2012wr012195. http://dx.doi.org/10.1029/2012WR012195.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges, 1994: A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res.*, **99**, 14415–14428, doi:10.1029/94jd00483. http://dx.doi.org/10.1029/94JD00483.

Maurer, E. P., A. W. Wood, J. C. Adam, D. P. Lettenmaier, and B. Nijssen, 2002: A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States. *J. Clim.*, **15**, 3237–3251, doi:10.1175/1520-0442(2002)015<3237:althbd>2.0.co;2. http://dx.doi.org/10.1175/1520-0442(2002)015%3C3237:ALTHBD%3E2.0.CO;2.

Mizukami, N., M. P. Clark, A. J. Newman, A. W. Wood, E. D. Gutmann, B. Nijssen, O. Rakovec, and L. Samaniego, 2017: Towards seamless large-domain parameter estimation for hydrologic models. *Water Resour. Res.*, doi:10.1002/2017WR020401. http://doi.wiley.com/10.1002/2017WR020401 (Accessed September 29, 2017).

Nash, J. E., and J. V Sutcliffe, 1970: River flow forecasting through conceptual models part I — A discussion of principles. *J. Hydrol.*, **10**, 282–290, doi:https://doi.org/10.1016/0022-1694(70)90255-6. http://www.sciencedirect.com/science/article/pii/0022169470902556.

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

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

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

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

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

Pushpalatha, R., C. Perrin, N. Le Moine, and V. Andréassian, 2012: A review of efficiency criteria suitable for evaluating low-flow simulations. *J. Hydrol.*, **420–421**, 171–182, doi:https://doi.org/10.1016/j.jhydrol.2011.11.055. http://www.sciencedirect.com/science/article/pii/S0022169411008407.

Rakovec, O., and Coauthors, 2016a: Multiscale and Multivariate Evaluation of Water Fluxes and States over European River Basins. *J. Hydrometeorol.*, **17**, 287–307, doi:doi:10.1175/JHM-D-15-0054.1. http://journals.ametsoc.org/doi/abs/10.1175/JHM-D-15-0054.1.

Rakovec, O., R. Kumar, S. Attinger, and L. Samaniego, 2016b: Improving the realism of hydrologic model functioning through multivariate parameter estimation. *Water Resour. Res.*, **52**, 7779–7792, doi:10.1002/2016wr019430. http://dx.doi.org/10.1002/2016WR019430.

Samaniego, L., R. Kumar, and S. Attinger, 2010: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resour. Res.*, **46**, W05523, doi:10.1029/2008wr007327. http://dx.doi.org/10.1029/2008WR007327.

Samaniego, L., and Coauthors, 2018: Anthropogenic warming exacerbates European soil moisture droughts. *Nat. Clim. Chang.*, **8**, 421–426, doi:10.1038/s41558-018-0138-5. https://doi.org/10.1038/s41558-018-0138-5.

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

Shafii, M., and B. A. Tolson, 2015: Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. *Water Resour. Res.*, **51**, 3796–3814, doi:10.1002/2014wr016520. http://dx.doi.org/10.1002/2014WR016520.

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

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

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

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

Westerberg, I. K., T. Wagener, G. Coxon, H. K. McMillan, A. Castellarin, A. Montanari, and J. Freer, 2016: Uncertainty in hydrological signatures for gauged and ungauged catchments. *Water Resour. Res.*, n/a-n/a, doi:10.1002/2015wr017635. http://dx.doi.org/10.1002/2015WR017635.

Wobus, C., and Coauthors, 2017: Modeled changes in 100 year Flood Risk and Asset Damages within Mapped Floodplains of the Contiguous United States. *Nat. Hazards Earth Syst. Sci.*, **2017**, 1–21, doi:10.5194/nhess-2017-152. https://www.nat-hazards-earth-syst-sci.net/17/2199/2017/nhess-17-2199-2017.html.

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

Yadav, M., T. Wagener, and H. Gupta, 2007: Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. *Adv. Water Resour.*, **30**, 1756–1774, doi:http://dx.doi.org/10.1016/j.advwatres.2007.01.005. http://www.sciencedirect.com/science/article/pii/S0309170807000140.

Yilmaz, K. K., H. V Gupta, and T. Wagener, 2008: A process-based diagnostic approach to model

evaluation: Application to the NWS distributed hydrologic model. *Water Resour. Res.*, **44**, W09417, doi:10.1029/2007wr006716. http://dx.doi.org/10.1029/2007WR006716.