Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-391
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 3 August 2018

Hydrology and Earth System Sciences
Discussions

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

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

[1]National Center For Atmospheric Research, Boulder CO
[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

Correspondence: Naoki 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 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, 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 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. Meanwhile, the use of KGE results in annual peak flow estimates that are better than from NSE, 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. Overall this work highlights the need for a fuller understanding of performance metric behavior and design in relation to the desired goals of model calibration.

1 Introduction

Computer-based hydrologic, land-surface, and water balance models are used extensively to generate continuous long-term hydrologic simulations in support of water 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).

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-391
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 3 August 2018

Hydrology and Earth System Sciences
Discussions

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; Pushpalatha et al., 2012; Price 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.

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; Yilmaz et al., 2008; Westerberg and McMillan, 2015; Westerberg et al., 2016; Addor et al., 2017a) has been gaining popularity for hydrologic model calibration (Yadav et al., 2007; Westerberg et al., 2011; Shafii and Tolson, 2015). A hydrologic signature is a statistic 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).

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.

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).

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?

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-391
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 3 August 2018

Hydrology and Earth System Sciences
Discussions

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?

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., 2013b) 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) lack of community-wide awareness of the pitfalls associated with use of squared error type metrics for high flow estimation.

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

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. 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., variability is not underestimated only in the ideal but unachievable situation when the correlation is 1.0. They further show that this results in a tendency to underestimate high flows while overestimating low flows (see Fig.3 in Gupta et al., 2009).

Our recent large sample calibration study (Mizukami et al., 2017) made us strongly aware of the practical implications of this problem associated with the use of NSE for model calibration. Figure 1 illustrates the bias in the model's ability to reproduce high flows when calibrated with NSE. The plot shows distributions of annual peak flow bias at 492 Hydro-Climate Data Network (HCDN) basins across the CONUS for the VIC model using 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

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-391
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 3 August 2018

Hydrology and Earth System Sciences
Discussions

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.

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:

$$KGE = 1 - \sqrt{[S_r(r-1)]^2 + [S_\alpha(\alpha-1)]^2 + [S_\beta(\beta-1)]^2} \qquad \alpha = \frac{\sigma_s}{\sigma_o}, \beta = \frac{\mu_s}{\mu_o} \tag{1}$$

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

The results of the Mizukami et al. (2017) large sample study motivated us to carry out further experiments to investigate how the choice of performance metric affects the estimation of peak 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 Datasets, Methods and Methods

We used 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 studies (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., 2016b) and is currently being used in application studies (e.g., Thober et al., 2018; Samaniego et al., 2018). Each of the models was independently calibrated to each of the 492 HCDN basins 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 (Newman et al., 2014; Addor et al., 2017b). The use of this large sample dataset helps to obtain more general and statistically robust conclusions (Gupta et al., 2014). We split the hydrometeorological data into a calibration period (October 1, 1999 - September 30, 2008) and an evaluation period (October 1, 1989 - September 30, 1999) and used a prior 10-year warm-up when computing the statistics for each period.

The model parameters calibrated for each model are the same as previously discussed: VIC (Newman et al., 2017; Mizukami et al., 2017) and mHM (Rakovec et al., 2016a, b). Although alternative calibration parameter sets have also been used by others,

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-391
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 3 August 2018

Hydrology and Earth System Sciences
Discussions

particularly for VIC (Newman et al., 2017), the purpose of this study is purely to examine the effect of performance metrics used for calibration, not to obtain "optimal" parameter sets. Each model was identically configured for each of the 492 basins, and 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.

Optimization was performed using the Dynamically Dimension Search (DDS, Tolson and Shoemaker, 2007). Five performance metrics were used: three based on KGE with varying scaling factors to emphasize different components, one being an application-specific high flow metric, and our benchmark performance metric 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.

Because variability is strongly correlated with annual peak-flow error (see Fig. 1 c), 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 results shown in this paper use KGE computed with S_r, S_α and S_β all set to 1.0.

For our application-specific high flow metric we use the Annual Peak Flow Bias (APFB) measure defined as:

$$APFB = \sqrt{[(\mu_{peakQ_s}/\mu_{peakQ_o} - 1)]^2} \quad (2)$$

where μ_{peakQ_s} is mean of simulated annual peak flow series and μ_{peakQ_O} is mean of observed annual peak flow series.

4 Results

4.1 Overall Simulation Performance

First, we focus on the general overall flow performance as measured by the performance metrics used. Figures 2 and 3 show the cumulative distributions of evaluation period model skill 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). Consider first the result obtained using KGE. For both models, at the median values of the distributions, use of KGE improves variability error by approximately 20% over NSE (Figs. 2a and 3a); however, the plots indicate a continued statistical tendency to underestimate observed flow variability even when the (1,5,1) component weighting is used. 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).

For both models, the use of APFB as calibration metric yields poorer performance 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 obtained when calibrating with NSE (plots not shown). This is consistent with

expectation, because improvement in variability error (α closer to unity) is known to cause a reduction in NSE optimality. In general, all the calibration results from both models are consistent with the NSE-based calibration characteristics described in Gupta et al. (2009).

4.2 High flow simulation performance

Next, we focus on the specific performance of the models in terms of simulation of high flows. As expected, use of the application-specific APFB metric (Eq. 2) leads to the best estimation of annual peak flows for both models (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 by approximately 5% for VIC and 10% for mHM at the median value during evaluation. 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 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.

Figure 4 c and d show the high flow simulation performance in terms of another high flow related metric - the %bias of the volume of above the 80 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 (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 results also exhibit larger variability in FHV bias across the study basins.

4.3 Implication for flood frequency estimation

Annual peak flow estimates are used directly 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 of annual flood magnitudes, we use the combined calibration and validation periods.

Figure 5 shows results that are consistent with Figure 4, although more outlier basins were 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 return periods for both models. mHM especially exhibits a clear reduction of the bias by 10% at the median compared to NSE. 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

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-391
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 3 August 2018

though the %bias of annual peak flow is less than 5% at median 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 with longer return periods.

5 Discussion

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.

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).

6 Conclusions

The use of large sample catchment calibrations of two different hydrologic models with five 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 annual peak flow 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. Calibration metric choice impacts high flow estimates very similarly for both models, although mHM provides overall better performance than VIC for all metrics evaluated.

2. 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.

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 models that are more complex, we can expect that other models would exhibit similar results when using KGE or a scaled variant. It seems clear that careful thought needs to be given to the design of application specific metrics if we are to obtain good performance for both the target metric (used for calibration) and other related metrics (used for evaluation), so as to increase confidence in the robustness and transferability of the calibrated model. This issue needs to be examined in more detail.

Code and data availability. Model calibration was performed using MPR-flex available at https://github.com/NCAR/mpr-flex/tree/direct_calib for VIC. mHM is calibrated with the MPR strategy implemented in the mHM http://www.ufz.de/index.php?en=40114. Hydrometeorological data are obtained from a part of Catchment Attributes and Meteorology for Large-sample Studies (CAMELS; Newman et al., 2014; Addor et al., 2017b). Analysis and plotting codes are available at https://github.com/nmizukami/calib4ffa/blob/master/ffa.ipynb.

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

Acknowledgements. This work was financially supported by the U.S Army Corps of Engineers Climate Preparedness and Resilience program.

References

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

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

Bergström, S.: The HBV model, in: Comput. Model. Watershed Hydrol., edited by Singh, V., chap. The HBV mo, Water Resouces Publications, Highlands Ranch Co., 1995.

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

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

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

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

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

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

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

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

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

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

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

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

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

Liang, X., Lettenmaier, D. P., Wood, E. F., and Burges, S. J.: A simple hydrologically based model of land surface water and energy fluxes for general circulation models, J. Geophys. Res., 99, 14 415–14 428, https://doi.org/10.1029/94jd00483, http://dx.doi.org/10.1029/94JD00483, 1994.

Maurer, E. P., Wood, A. W., Adam, J. C., Lettenmaier, D. P., and Nijssen, B.: A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States, J. Clim., 15, 3237–3251, https://doi.org/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, 2002.

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

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

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

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

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

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

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

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

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

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

Samaniego, L., Kumar, R., and Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, Water Resour. Res., 46, W05 523, https://doi.org/10.1029/2008wr007327, http://dx.doi.org/10.1029/2008WR007327, 2010.

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

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

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

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

Thober, S., Kumar, R., Wanders, N., Marx, A., Pan, M., Rakovec, O., Samaniego, L., Sheffield, J., Wood, E. F., and Zink, M.: Multi-model ensemble projections of European river floods and high flows at 1.5, 2, and 3 degrees global warming, Environ. Res. Lett., 13, 14 003, http://stacks.iop.org/1748-9326/13/i=1/a=014003, 2018.

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

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

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

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

Wobus, C., Gutmann, E., Jones, R., Rissing, M., Mizukami, N., Lorie, M., Mahoney, H., Wood, A. W., Mills, D., and Martinich, J.: 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, https://doi.org/10.5194/nhess-2017-152, https://www.nat-hazards-earth-syst-sci.net/17/2199/2017/nhess-17-2199-2017.html, 2017.

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

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

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resour. Res., 44, W09 417, https://doi.org/10.1029/2007wr006716, http://dx.doi.org/10.1029/2007WR006716, 2008.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-391
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 3 August 2018

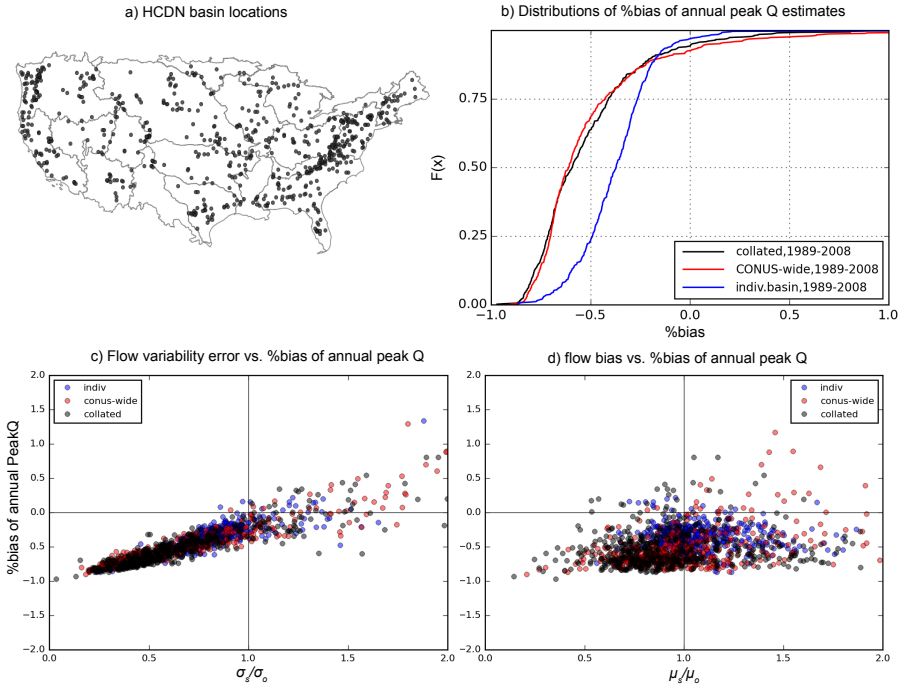

Figure 1. a) Spatial distribution of Hydro-Climate Data Network (HCDN) basins, b) Cumulative distribution of %bias of annual peak flow over 1989-2008 simulated with three different sets of VIC parameters used in Mizukami et al. (2017) at HCDN basins. c) Relationships between variability error (simulation to observation ratio of daily flow variability) with %bias of annual peak flow. D) Relationships between mean error (simulation to observation ratio of mean flow) with %bias of annual peak flow

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-391
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 3 August 2018

Hydrology and Earth System Sciences
Discussions

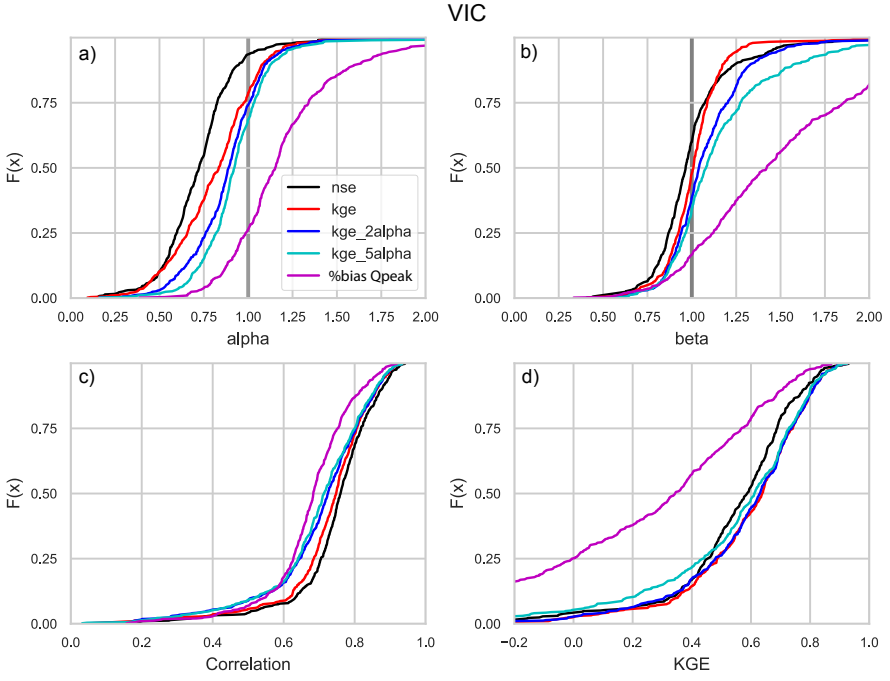

Figure 2. Cumulative distributions of a) flow variability errors α, b) bias β, c) linear correlation r, and d) Kling-Gupta Efficiency over 492 HCDN basin calibrations with 5 objective functions. Metrics are based on the simulation during the validation period.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-391
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 3 August 2018

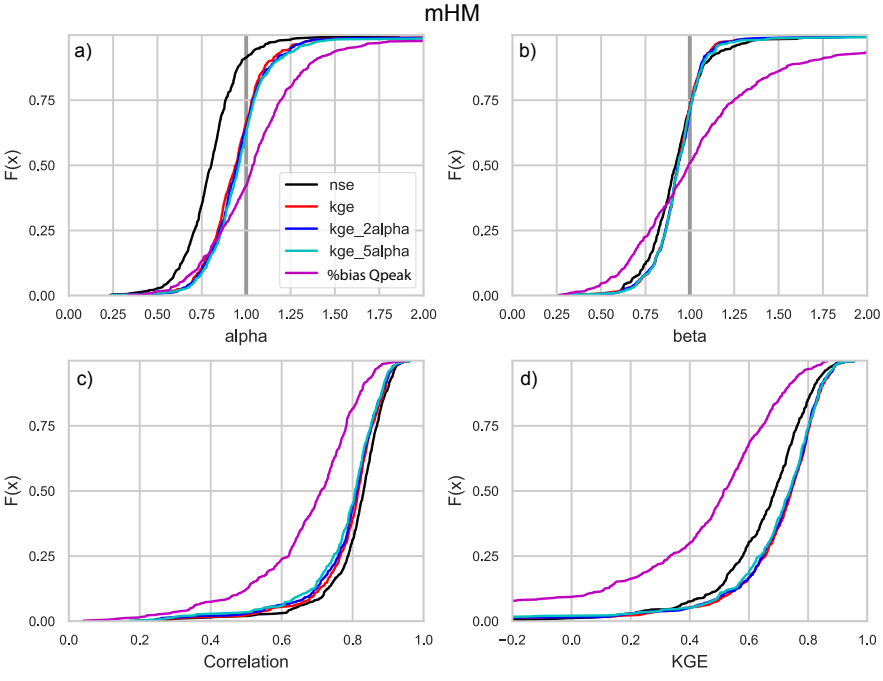

Figure 3. The same as Figure 2 except for mHM.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-391
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 3 August 2018

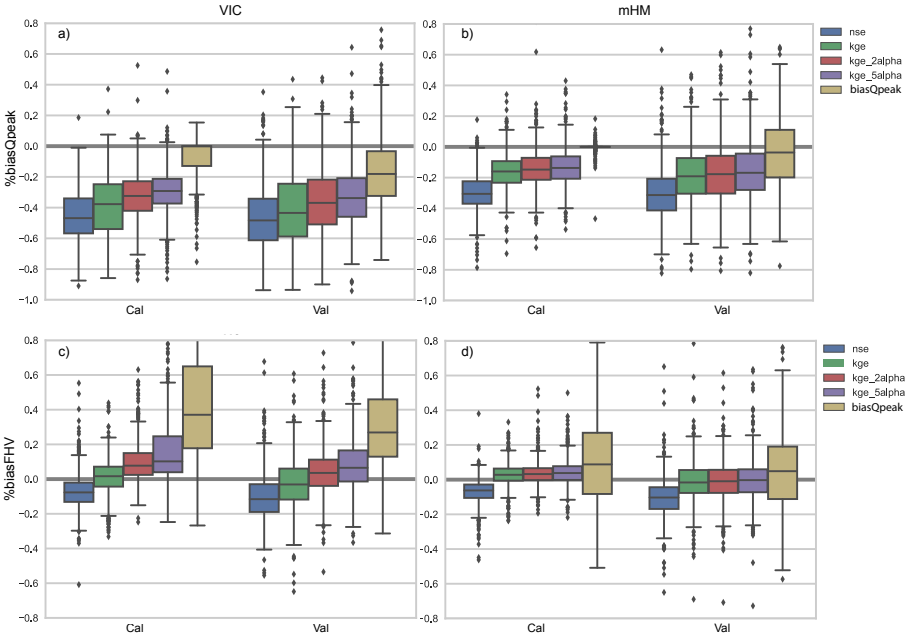

Figure 4. Boxplots of percentage bias of annual peak flow (top row) and flow volume above 80 percentile flow duration curve (bottom row) over the 492 HCDN basin calibrations with 5 objective functions for calibration and validation periods and two models.

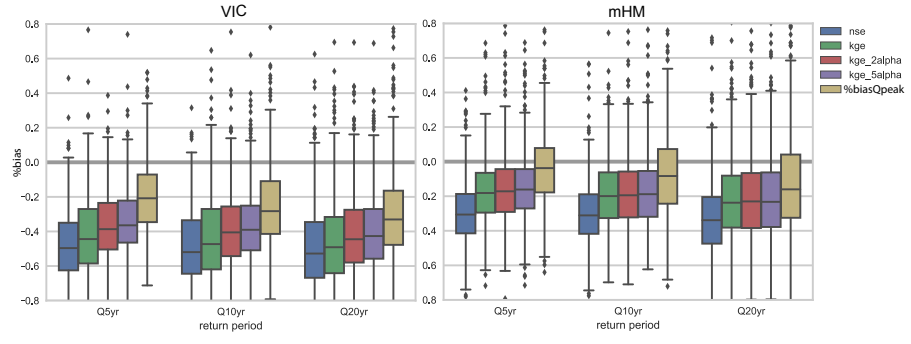

Figure 5. Boxplots of absolute error of flood estimates corresponding to three return periods (5-yr, 10-yr and 20-yr) over the 492 HCDN basins for the two models.

Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-391
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Discussion started: 3 August 2018

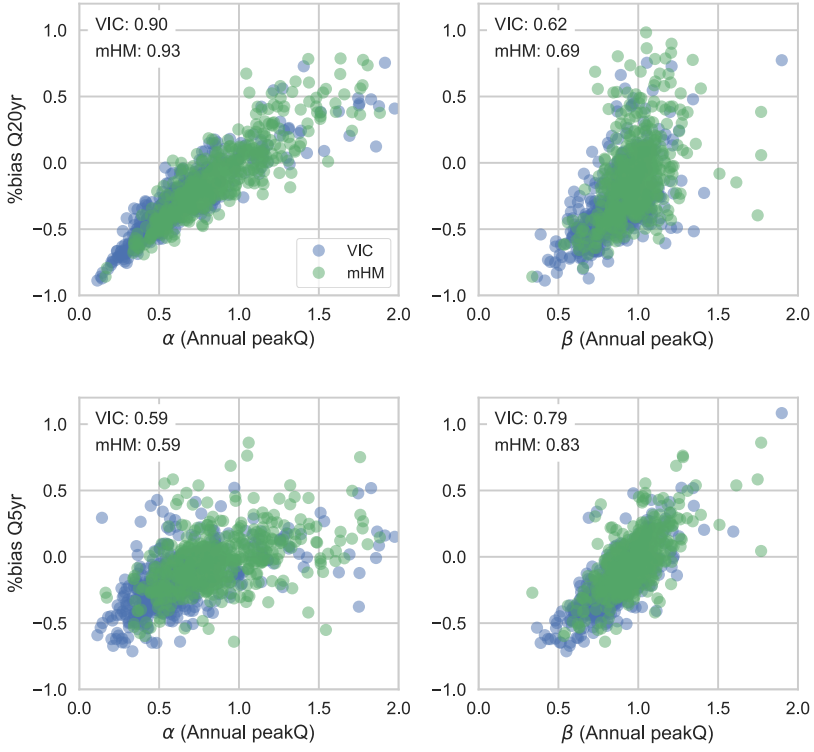

Figure 6. Scatter-plots between a) simulation-observation ratio of variability of annual peak flow series (α) and %bias of 20-yr flood magnitude, b) simulation-observation ratio of mean annual peak flow series (β) and %bias of 20-yr flood magnitude, c) and d) are the same as a) and b) except for 5-yr flood magnitudes. Linear correlations between two variables are specified at upper-left corner of each plot.