Reply to anonymous referee 1

Here we provide our responses to comments from referee #1, including the plan to revise the manuscript in response to reviewer comments. The original review comments are in black, and our responses in red.

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 really appreciate these thoughtful comments. We completely agree with the main comment – that sum-of-squared error metrics commonly used in optimization reduce the variance (by design), and hence representing extremes requires stochastically simulating the error term. This is in fact a key component of our related work on probabilistic quantitative precipitation estimation (Newman et al., 2015).

We carefully reviewed the paper by Famer and Vogel (2016) and also Montanari and Brath (2004). Both discuss about the stochastic estimation of model errors. The modeling approach we took is deterministic, but designed to examine how high flow statistics are sensitive to different evaluation metrics. We are interested in analyzing the distributions of errors for models calibrated using different objective functions, as well as examining high flow bias for experiments where we stochastically generate residual errors in deterministic calibrated VIC/mHM simulations. This analysis can be done with the existing simulations used for this manuscript, and, as such, we do not anticipate much time-consuming analysis. We plan to add the results from this additional analysis in new section. We will also revise the conclusions by incorporating the new results in the revised manuscript. We will summarize both results; error characteristics from KGE/NSE calibrated flow and ensemble flows based on stochastically generated error added to KGE/NSE based streamflow simulations.

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.

The ideas illustrated in the Farmer and Vogel (2016) paper will help us better frame our contribution. While we arrived at similar conclusions to Farmer and Vogel (2016) via different means, we feel that these issues are still not well appreciated by the broader community. We feel that our paper provides additional explanations of unintended consequences of model calibration decisions.

Reference

Farmer, William H., and Richard M. Vogel. On the deterministic and stochastic use of hydrologic models. Water Resources Research 52.7 (2016): 5619-5633.

Montanari, A., and A. Brath (2004), A stochastic approach for assessing the uncertainty of rainfall-runoff simulations, Water Resour. Res., 40, W01106, doi:10.1029/2003WR002540.

Newman, A.J., M.P. Clark, J. Craig, B. Nijssen, A. Wood, E. Gutmann, N. Mizukami, L. Brekke, and J.R. Arnold, 2015: Gridded Ensemble Precipitation and Temperature Estimates for the Contiguous United States. *J. Hydrometeor.,* **16**, 2481–2500, https://doi.org/10.1175/JHM-D-15-0026.1