

Responsive comments to Referee #3

The Referee #3 has two major concerns about the rationality and methodology of the study.

1. The practical applicability of such an approach is highly questionable.

Automatic calibration is important for practical application of hydrological models. The opinion that a **single-objective function (SOF)** cannot capture all of the important characteristics of the observed data has been gradually accepted. More and more hydrologists seek to improve the calibration methods to capture various aspects of hydrologic responses simultaneously. In this study, we hypothesized that SOF is still worthy studying if properly defined. Therefore, **the purpose of our study** is to demonstrate the potential capability of SOF to simultaneously address multi-response modes of the hydrograph. By this work, the authors are not intended to oppose the current researches on multi-objective calibration. We are not intended to advocate the practical applicability of the proposed optimized SOF, either. What we want to do is to remind the community that we may not exploit the full capacity of SOF for model calibration. By this work we show that the popular opinion of our community may not be necessarily right that no SOF can represent all the important characteristics of even one specific kind of hydrological variable (e.g., streamflow).

I acknowledge that we do not provide a practical way to apply our approach in model application. Although, our study still provides valuable insights for the community.

2. The methodology is not technically sound. Why apply a single-objective function that is tuned using an aggregate of four objective functions? One could rather apply the aggregated single-objective function directly for the optimisation. Using an aggregate of individual objective functions would be more transparent in relation to which hydrograph behaviours are balanced in

the optimisation.

I believe that the Referee refereed to the composite likelihood index (CL) by “an aggregate of individual objective functions”. This question is similar to Comment #1 by Referee #1. To the best of authors’ knowledge, this aggregated index cannot be used directly in the model calibration. From the definitions of CL below, we can see that all the individual objective functions (i.e., *NSE*, *TRMSE*, *ROCE*, *SFDCE*) are required to be available in order to calculate the corresponding max/min values (for the purpose of normalization). To be noted, *L* in the equation is the total number of calibration runs. That means CL can be used to make a comparison evaluation on existing model parameter sets, it cannot be used to optimize model parameter set. In the original literature proposing CL (Price et al., 2012), the authors concluded that the CL calibration showed promising performance in model validation—greater than NSE—which encourages further use of this approach for **scenario-based predictive modeling**. This also indicates the potential usage of CL is not for calibration but for scenario-based predictive modeling (in which the model runs are determined in advance).

$$\begin{aligned} NSE &= \frac{\max(0, NSE_i)}{\sum_{i=1}^L \max(0, NSE_i)} \\ TRMSE &= \frac{1 - \min(1, |TRMSE_i|)}{\sum_{i=1}^L 1 - \min(1, |TRMSE_i|)} \\ ROCE &= \frac{1 - \min(1, |ROCE_i|)}{\sum_{i=1}^L 1 - \min(1, |ROCE_i|)} \\ SFDCE &= \frac{1 - \min(1, |SFDCE_i|)}{\sum_{i=1}^L 1 - \min(1, |SFDCE_i|)} \end{aligned}$$

In our study, we used CL to evaluate the performance of model runs calibrated by the proposed SOF with different exponents. The best exponent can be determined by such evaluation. We acknowledge that such evaluation is not necessarily sufficient. Theoretically, the best way to make the evaluation is expert inspection (i.e., Turing test, https://en.wikipedia.org/wiki/Turing_test). For

a large number of case studies (196 MOPEX watersheds), the expert inspection is not practical and, therefore, the objective index CL was used to aid in the selection procedure.

I would like to focus on the aggregated single-objective function concern specifically raised by the Referee. I believe that the whole methodology can be understood with clarification of this issue.