Responsive comments to Referee #1

1. *CL is another single objective function (it is not a multi-objective function). Therefore, the first question that arises is why not using CL directly for model calibration?*

   The similar comment has also been raised by the Referee #3. 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, but 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).

$$_{NSE} = \frac{\max(0, NSE_i)}{\overset{L}{\underset{i=1}{}} \max(0, NSE_i)}$$

$$_{TRMSE} = \frac{1 \quad \min(1, |1 \quad TRMSE_i|)}{\overset{L}{\underset{i=1}{}} 1 \quad \min(1, |1 \quad TRMSE_i|)}$$

$$_{ROCE} = \frac{1 \quad \min(1, |1 \quad ROCE_i|)}{\overset{L}{\underset{i=1}{}} 1 \quad \min(1, |1 \quad ROCE_i|)}$$

$$_{SFDCE} = \frac{1 \quad \min(1, |1 \quad SFDCE_i|)}{\overset{L}{\underset{i=1}{}} 1 \quad \min(1, |1 \quad SFDCE_i|)}$$

2. *In a second experiment, the Authors compare the simulations obtained by calibrating on the Nash Sutcliffe, to the ones obtained with calibration on F1.*

The comparison is done by evaluating model performance on the individual components of the function CL. However, because the exponent of the function F1 has been chosen as to optimize CL, one can say that CL has been used as an objective of model calibration. The comparison is unfair, as it is quite obvious that the model performance will be better with respect to an objective function to which the model has been optimized than with respect to an objective function to which the model has not been optimized.

The purpose of our study is to demonstrate the potential capability of single objective function to simultaneously address multi-response modes of the hydrograph. Under this umbrella, the purpose of the second experiment is to demonstrate that there does exist a single-objective function that can compromise multi-response modes of hydrograph, which is usually not the traditionally used single-objective function NSE. For this purpose, our comparison by individual components of the function CL can support our conclusion. Maybe, our expression 'to verify the advantage of the proposed objective function' leads to the confusion, which will be modified in the revised manuscript.

As we discussed in the reply to first comment, the CL index cannot be used for model calibration. This could be a misunderstanding by the Referee. To the authors' understanding, the Referee implies that calibration on CL (an aggregate of individual functions) would definitely result in the sound performance of hydrograph in terms of multi-response modes. However, this is not necessarily true even if CL could be used for calibration. Otherwise, it is so easy to solve model calibration issue by combining different objective functions into one aggregate index.

3. *In a third experiment, the Authors compare the simulation obtained with respect to F1 to the envelope of curves obtained by multi-objective calibration on the individual components of CL. They show that their optimal model lies*

within the envelope. Again, because the exponent of their function has been chosen to optimize CL, their simulation would lie very close to the Pareto front of the 4 objectives of CL. Again, this is not a strong test for assessing the quality of their objective function.

Please refer to the response to the second comment.

4. *In a fourth experiment, the Authors perform time validation and show model performance on FDCs. This is, in my opinion, the only valid and independent test performed in this study, and it shows that the observed FDC can be very distant from the simulations. This test does not really support the conclusions of this study.*

   This comment is also related to the purpose of our experiment. I thank the Referee for the only positive words for our work, and I want to re-state that the purpose of this fourth experiment is to demonstrate the competence of a single objective calibration compared to multi-objective calibration. We are not meant to argue that our proposed single objection function can do perfect work under all circumstances. The Figures 6 and 7 can well support our conclusion that “single-objective calibration with the OSOF can compromise multi-response modes of the hydrograph to obtain a relatively sound simulation, which is comparable to the result of multi-objective calibration”.

5. *I think the Authors should provide other means to identify the parameter alfa in F1. For example, by choosing a value that minimizes the heteroscedasticity of the residuals.*

   I think the above discussions can address this comment. We acknowledge that there exist many other evaluation criteria that can be chosen, including the suggested heteroscedasticity of the residuals. In fact, referring to Turing test (https://en.wikipedia.org/wiki/Turing_test), we can say that the final solution to evaluate the model performance is expert inspection. Any choice of evaluation criteria will have its bias under different circumstances. The four evaluation criteria (*NSE, TRMSE, ROCE, SFDCE*) are four widely used objective functions focusing on peak flows, low flows, water balance, and flashiness, respectively. It is reasonable to make such choice at the beginning of such studies.

6. *The Authors should also look at the problem of estimating lambda in a Box Cox transformation, as this is very much related to their problem. Standard least squares lead to a maximum likelihood estimator that is essentially function F1 with exponent 2 (or equivalently, the Nash Sutcliffe efficiency). A box cox transformation results in a slightly different likelihood, which depends on lambda, and there are different ways of estimating lambda, including that of maximising the likelihood.*

   I seriously think this question is out of scope of our study.

7. *The comparison of different objective functions should use widely used metrics, it is ok to use NS and CL, with inclusion of some other objective functions previously proposed, such as the Kling Gupta.*

   We can do that in a further study. We can compare even more metrics, e.g., the 19 objective functions mentioned in Price et al. (2012), but it is a totally different work if the Referee can acknowledge the purpose of this study.

   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., 48, W10542, doi:10.1029/2012WR012005.

8. *The evaluation should be done with metrics independent on the objective functions used, such as FDCs, QQ plots, indices that are deemed important by hydrologists such as the baseflow and flashiness index, and preferably in the calibration period.*

   I think this concern can be addressed from all the above discussions.