Responsive comments to Referee #2

We thank the Referee for the general positive comments for our work. The following are the responsive comments point-by-point.

1. *The authors need to clarify their approach in dealing with model uncertainty. Whatever objective function is used, a single-objective calibration method always produces a single parameter set. Consequently, it cannot quantify model uncertainty unless it is placed within an uncertainty framework (e.g. Bayesian methods) where residual error modelling is part of the model calibration. The authors haven't mentioned such a framework in their paper, as a result their single-objective calibration method has no ability to quantify model uncertainty and cannot be compared with a multi-objective calibration method from the aspect of estimating uncertainty bounds. We suggest removing all reference to model uncertainty in the paper, especially in Section 4.4 and Figure 7.*

Reply: Thanks for the comments. We will revise the manuscript accordingly.

2. *The conclusions of the paper rely heavily on Figure 2, 5 and 6. However, the figures are hard to read and aggregate a lot of information, which makes it difficult to judge on the validity of the associated comments. We suggest adding several tables or figures to clarify the findings. For example:*
    1) *Figure 2: The impact of the two metrics ROCE and SFDCE is difficult to distangle from the two other metrics (NSE and TRMSE). The use of the size and color of individual dots is not recommended for such a dense plot. We suggest using standard 2d scatter plots showing the relationships between two metrics only.*
    2) *Figure 5 and 6: Hydrograph plots are generally noisy and hard to comment on, unless they are plotted over a very short time scale (e.g. a flood event). We suggest redrawing the plots using flow duration curves*

(similar to figure 7) and 2 or 3 flood events. In addition, the analysis of hydrographs on 8 sites is not sufficient to judge on the quality of model calibration method. If the intend of the authors is to show the similarity between the OEV calibration and an alternative method, we suggest computing a similarity metric between simulations produced from the two methods. For example the four objective functions NSE, TRMSE, ROCE and SDFCE can be used where Qo,t is replaced with the simulated value produced with the OEV calibration. Low values of these metrics would suggest that both simulations are similar. This approach would offer a quantitative approach in the comparison of simulations.

Reply: Thanks for the comments. For Figure 2, ROCE and SFDCE are indeed difficult to be recognized. We want to show the four metrics in one plot so the readers can get a holistic view of the evaluation metrics. As we discussed in the manuscript, there is no obvious trend detected from the figures in terms of ROCE and SFDCE. The general performances of ROCE and SFDCE are reasonably well for most calibrations in this study. It is the reason we did not focus on these two metrics. We can add a new 2d scatter plot to show them more clearly.

For Figure 5 and 6, we can follow the suggestions to plot the FDC curves. Also, we can calculate the metrics between simulations produced from the two methods, and list them in Table 2.

3. *We believe that one of the reasons behind the similarity between the single and multi-objective calibration results reported by the authors comes from the lack of diversity in the objective functions selected for the multi-objective calibration exercise. More specifically, the NSE and TRMSE are both metrics that compute the sum of squared residuals of flow simulations (see detailed comment #1). TRMSE uses an Box-Cox transform with exponent 0.3, which is not putting a very strong emphasis on low flows. As suggested by Pushpalatha*

et al. (2012), an exponent between −1 and 0 (log function) would be more appropriate. Such an exponent would clearly distinguish a calibration based on NSE, which focuses on high flows as indicated by the authors, from a calibration based on TRMSE.

Reply: We can add the relevant discussions in the manuscript.

4. *In their conclusion, the authors claim that "the methodology was applied to 196 (...) watersheds" (page 25, line 387). Such a large scale testing provides a strong support for the authors' conclusions. However, we noted two important shortcomings in the way the authors used the MOPEX catchment dataset:*
    1) *First, the authors reduced the catchment data set from an initial list of 438 catchments to 196. The authors indicate that the catchments were selected "because of the applicability of the Xinanjiang model". This point is a major problem because we believe that a calibration method should not be tested on good behaving catchments only. Trade-off in calibration between different objective functions appear when the model cannot reproduce the full extent of the flow regime, which is generally a synonym of poor model performance. As a result, we suggest expanding the dataset used in this paper to catchments where the model does not perform well, and check if the paper conclusions still hold in less favourable modelling conditions.*
    2) *Second, the authors tested their calibration method on 196 catchments, but only reported validation statistics on 8 "representative catchments" (see page 17, line 305) that were selected "arbitrarily" within each OEV group (see Page 16, Line 287). The authors do not provide additional details on the rationale behind this selection process. We believe that this point constitutes a major issue in the paper, where important conclusions are drawn from a very small sub-sample of the*

initial dataset. We strongly recommend reporting the validation results on the full set of 196 catchments, or if possible, on the complete MOPEX dataset.

Reply: 1) the selection of 196 MOPEX watersheds. The selection of study catchments is necessary to try to avoid (although never completely) the tangle of model uncertainty and parameter uncertainty. In this study, our focus is the objective function, which is related to parameter uncertainty. If model structure error is large, the conclusions would have large uncertainties too. Of course, the impact of model structure uncertainty on the objective function is another interesting question, which can be explored by expanding the dataset used in this paper but can be another separate study.

2) Thanks for the comments. We will provide the evaluation indices for all 169 study catchments in Table 2.

5. *Detailed comments*

Reply: We will revise the equation and text accordingly.