Reply to anonymous referee 2

Here we provide our responses to comments from referee 2, including the plan to revise the manuscript in response to reviewer comments. The original review comments are in black, and our responses in red.

RC2

1 OVERALL RECOMMENDATION

The manuscript addresses the important topic of the choice of calibration metrics (CM) to be used for rainfall-runoff modeling, and presents results obtained on 492 US catchments. I found the paper interesting, including relevant references. If the presented results are not highly original, the paper is, in his present form, an excellent illustration of the limitation of the use of Nash and Sutcliffe efficiency metric (NSE, 1970) for model calibration. Nevertheless, I do have major comments on the used dataset, the applied methodology and the discussion part. Thus, I recommend to accept the manuscript in HESS with major revisions detailed below.

We thank the referee for taking time to review our manuscript. Yes, the purpose of the paper is to illustrate pitfall of NSE metrics for the model calibration using large-sampling basin context, motivated by the fact that the community (esp. for practical hydrologic application) keeps using NSE or squared mean error based metric for the model calibration without thinking much on its implications, though we agree this has been discussed in the previous paper (Gupta et al., 2008).

2. GENERAL COMMENTS

2.1 Description of the studied catchments

Even though the objective of such “large-sample hydrology paper” is not to present results obtained on a limited number of catchments having the same hydrometeorological characteristics but to have general conclusions on rainfall-runoff modeling, I think the diversity of the studied catchments has to be addressed and quantified.

This description is lacking right now in the paper. A presentation of the general characteristics of the studied catchments should be added in the paper, in order to understand the variability of catchments characteristics (catchment area, runoff coefficient, mean annual solid precipitation, etc.), especially in the context of flood modeling: what are different flood processes and dynamics included within this catchments sets (flash floods, snowmelt floods, rain-on-snow floods, groundwater floods, etc.)? Moreover, the timestep considered in the two rainfall-runoff models is not stated in the paper and should be mentioned. Are the models working at daily timestep? Is this timestep consistent with the flood dynamics of every studied catchment?

Yes, one of main objectives of large sample basin study is to generalize the conclusions drawn regarding hydrologic modeling evaluations (Gupta et al 2014). This manuscript used a subset of the catchments presented by Addor et al., (2018), who describe in detail the variability of climate/geophysical/hydrologic characteristics for the 672 US catchments. We have avoided repetitive description, but agree that it is good idea to include catchment variability closely related to “high flow” based on “the catchments used in this paper” to show that we cover a sufficient range of catchment variability. Also, some descriptions of spatial pattern of the model skills in addition to the distributions will be added in the revised manuscript.

2.2 Split-sample test

For every catchment, the calibration and validation periods are the same time-periods, 1999-2008 and 1989-1999, respectively. I think that performing a basic split and sample test (Klemeš, 1986) on each catchment would be particularly interesting in this context, especially to address temporal (in)stability of parameter sets obtained with particular CM (topic partially addressed page 6, line 12).

It may not be clear enough in the text but we did use a split-sample method i.e., calibration period during 1999-2008 and validation during 1989-1999 (Page 4, Line 29-31). Exception is that evaluation of annual peak flow error where we combined the two periods (Page 6, Line 24) to have increased sample size. We will attempt to describe this more clearly in the revised manuscript.

2.3 List of the studied CM

The paragraph listing the studied CM (page 5, lines 5 to 16) is unclear and would be easier to understand if a list (or table) of the five studied CM was added.

We will consider adding a table to list the metrics used in the revised manuscript.

2.4 “Application-specific” or “hydrologic signature”?

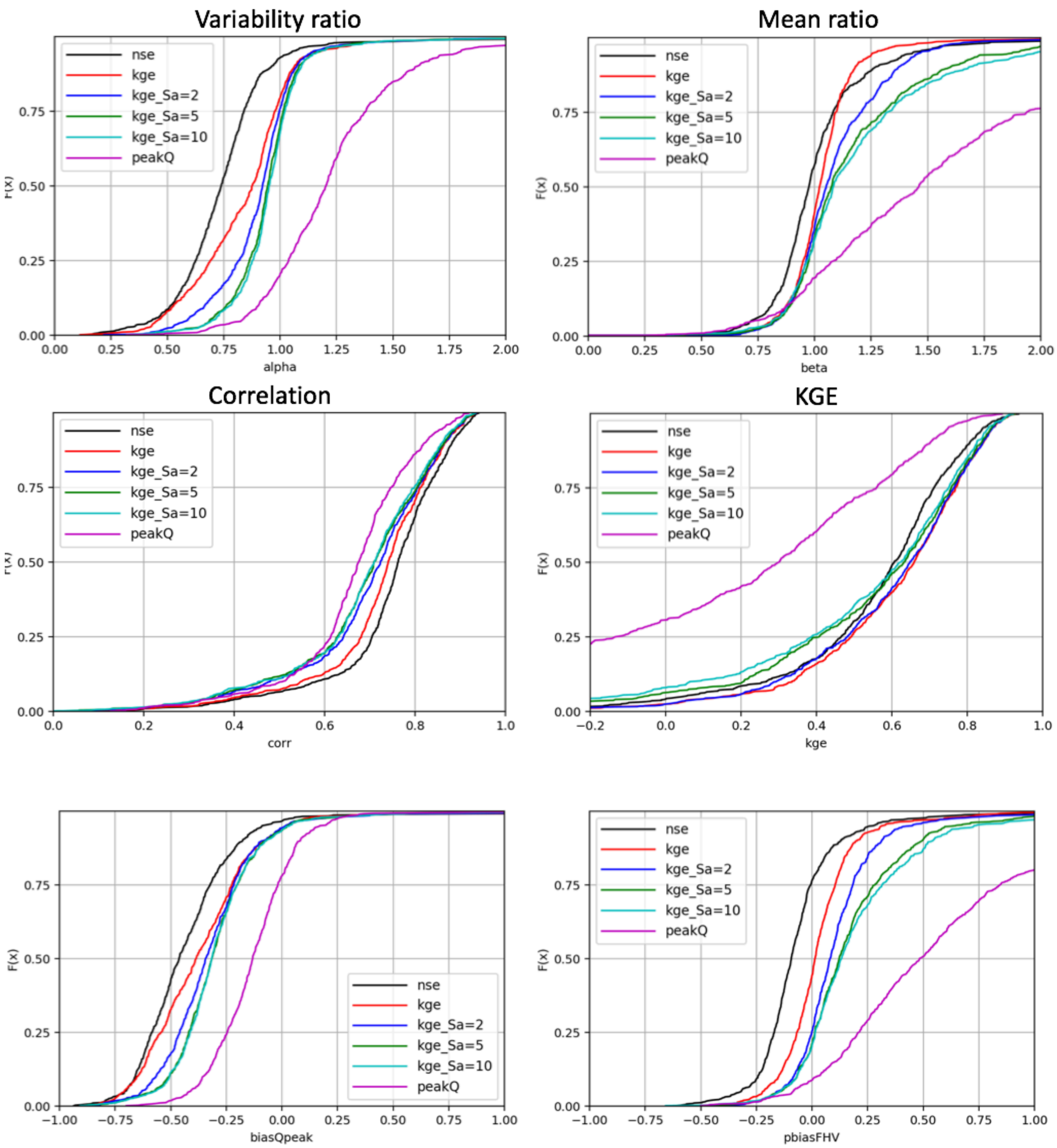
From page 5 to the end of the paper, APFB is named as an “application-specific” metric, while being introduced as an “hydrologic signature” (see definition of “hydrologic signature” in the paper introduction, page 2 lines 10 to 24) in the paper objective presentation. What is the difference between an “application specific” and a “hydrologic signature” CM in this context? Finally, is APFB an “application specific” or an “hydrologic signature” CM? Could you address this point?

The “hydrologic signature” is a metric derived from hydrologic variable time series to imply how a specific hydrologic process behaves, as expressed by the model simulation (or observation) for better understanding of the processes. So, it is process-oriented metric. On the other hand, while the “application-specific metric” can be a hydrologic signature, it is a metric used by water managers for specific decision-making. Therefore, it can be simple bias of monthly or annual time series, which can be used for long term hydroclimate assessment purpose (not worry about hourly, or daily variability). We will clarify the difference in the definition between “hydrologic signature” and “application-specific metric” in the revised manuscript.

2.5 Impact of the KGE scaling factors

The limited impact of the different KGE scaling factors used in the paper is very little discussed, while being particularly interesting. This point has to be discussed in the paper. Moreover, why not trying another combination with a larger variability ratio scaling factor, such as (Sr=1, Saplha=20, Sbeta=1), to assess a potential significant improvement of annual peak flow bias? What about another test with (Sr=0.1, Saplha=1, Sbeta=0.1) or even (Sr=0, Saplha=1, Sbeta=0)?

We actually tested a few extreme cases: (Sr=1, Saplha=10, Sbeta=1) and (Sr=0, Saplha=1, Sbeta=0). It did not show improvement of annual peak flow. These results (all the statistics) did not change significantly (statistically at 95%) from the case (Sr=1, Saplha=5, Sbeta=1) as shown below.



2.6 Figures

In general, the presentation of the figures could be improved for a better understanding:

- The performance metrics plotted have different names in the axis labels and in the figure legends (e.g. it is not explicit that “%biasFHV” is equal to “percentage bias of flow volume above 80 percentile flow duration curve” in the Figure 4) ;

- The name/typography of several performance metrics is changing over figures (“%bias Qpeak” on Figure 2 but “%biasQpeak” on Figure 4) ; - Why not using Greek letters in Figure 2 x-axis?

- The link between the five CM and the figure legend is never clearly stated, and for example, the reader has to guess that “kge_2alpha” is equal to (Sr=1, Saplha=2, Sbeta=1).

- How boxplots have been constructed? What are the outlier points plotted below and over the boxplots?

We will improve the figures based on the above comments and will also add the definition of boxplot construction in the revised manuscript.

3 SPECIFIC COMMENTS

1. Page 4, line 27: please change (Maurel et al., 2002) into Maurel et al. (2002).
2. Page 13, figure 2: please state in the figure legend that results presented in this figure are obtained with the VIC model.

We will incorporate the above two comments in the revised manuscript.

4 REFERENCES

Klemeš, V., 1986. Operational testing of hydrological simulation models. Hydrological Sciences Journal 31, 13. https://doi.org/10.1080/02626668609491024.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I – A discussion of principles. Journal of Hydrology 10, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.