Response to Review of HESSD-11-5599-2014 by Anonymous Reviewer #2

Note that original reviewer comments are in blue and author responses are in black throughout.

The article presents a new large datasets of catchments in the US built for hydrological modelling applications. The authors shortly present the dataset and then the application of the SAC-SMA model considered as a benchmark. The issues of model spatial variability of model performance and the weight of major model errors in overall performance criteria are discussed. This is a very valuable contribution, which should encourage the application of models on large datasets for various purposes (validation, regionalization, etc.). The article is generally clear and easy to follow. I have however four main suggestions to improve its content:

We thank the reviewer for their very thorough and thoughtful review. We have added figures and text in many places to address the reviewer's concerns.

A. The introduction should better review and acknowledge the past efforts to gather large datasets for hydrological applications in the US (e.g. the MOPEX dataset among others; see also the review of Gupta et al. (2014) in their supplementary material) and better explain to which extent this new dataset offers new opportunities for model testing compared to existing US datasets.
B. Although the main objective of the paper is to present this new dataset and the benchmark model application, the introduction could raise the scientific questions that the authors wish to specifically investigate in this article, e.g. related to the issues discussion in section 4.

We have reworked the introduction to include more cites from Gupta et al. (2014) and several mentions of MOPEX. We have also included discussion of how this dataset compares to and extends the MOPEX dataset. We have also included in the discussion reference to the scientific questions this paper is aiming to address.

C. The presentation of the data set could be improved, by introducing a more detailed description of the catchment physical characteristics.
We agree with this comment, it was an oversight to not include a more detailed description of the basin set in the paper. We have added two new figures along with panels in other figures to address this comment. We will also be including the basin descriptor data in the downloadable dataset.

D. I think the choice of the authors to use the classical Nash-Sutcliffe efficiency index as objective function for calibrating the benchmark model is questionable, given the clear deficiencies of this criterion, as demonstrated by the work of Gupta et al. (2009). I think this makes the proposed benchmark a bit outdated. It is now five years that Gupta et al. (2009) proposed their KGE criterion, and this paper is a good way to encourage the future users of the dataset to use more up-to-date approaches for model calibration and give up old habits that are clearly less efficient. Therefore I encourage the authors to present their results using the KGE criterion as objective function instead of RMSE-based criteria.

We thank the reviewer for their thoughtful consideration of the many issues regarding the choice of objective function. While we agree that KGE is likely a "better" objective function than RMSE and that RMSE is outdated, however KGE is essentially a reweighting of RMSE and still subject many of the same issues as RMSE, although to a lesser degree. We do provide the decomposition of RMSE into the three KGE components in Fig. 7.

We feel that having discussion on the limits of using RMSE as the objective function and presenting the results in terms of NSE while including the decomposition terms and discussion are worthwhile to the community. It highlights how RMSE performs for the various decomposition flow metrics and again highlights the need to use more innovative and thoughtful objective functions.

This study is intended to provide a benchmark dataset which provides the "old habit" approaches as the benchmark to advance forward and apply, say, KGE as the objective function, then compare to this benchmark for any type of streamflow based performance metric. As Reviewer #1 notes in their additional comment, this dataset allows for advanced calibration methodology experiments, one of which would be to use a more advanced objective function.

Minor Comments:

1. Section 2: The authors should give illustrations of the physical and hydroclimatic characteristics of the selected catchments. Distributions of catchment size, mean elevation, slope, or other descriptors, as well as basic hydroclimatic values (mean Q, P and PE), could give a better idea of the types of selected catchments.
We have added two figures (Figs. 2 & 3) along with two new paragraphs of discussion in the text. We will also include these data in the dataset.

2. Section 2: All readers may not be equally familiar with the geography of the US. Therefore, to better follow the discussions presented in section 4 on the spatial distribution of results, which refers to several specific regions or locations in the US, it might be useful to have a map (e.g. the map in Fig 1.a) that show these regions.
We have changed figure 1 to include references to specific geographic areas and changed the text to help clarify regional descriptions.

3. Section 2.1: A few lines could be added on data quality and availability. Are there indexes to qualify the reliability of streamflow data? What is the range of percentages of gaps in the series?
We have added some discussion of streamflow data quality control flags to the text. We are also including the available flags with the downloadable dataset

4. P. 5603, L. 23: write "contiguous United States (CONUS)"
We have spelled out this abbreviation

5. P. 5603, L. 12: What "MT-CLIM" stands for?
MT-CLIM stands for Mountain Climate simulator (MT-CLIM).

6. Section 3.1: The authors could shortly comment the existing past applications of this model on large datasets, especially in the US. What were the results? What is already known on the possible model limits across the US?
We have tried to find a few applications of this and similar models across CONUS. We have found little to go on. If the reviewer has a specific reference in mind, we would be glad to include it.

7. P. 5606, L. 19-21: By calibrating the model on the first half of the series and validating it on the second half, the authors only applied half of the Klemeš split-sample test (Klemeš, 1986). It would be useful to also do the reverse test, by calibrating the model on the second half and validating it on the first half. This would provide a benchmark simulation in validation mode on all available data (not only half of them) and hence a more comprehensive evaluation of model performance. This would also make the comparison of the difference in model performance between calibration and validation more interesting: by comparing the mean performance in validation on the two periods with the mean performance in calibration on the two periods, one avoid the possible bias resulting from the fact that the two periods may not be similarly difficult/easy to simulate. Last this would give the opportunity to comment the stability of parameter values between the two calibration periods and hence possibly identify regions where model parameter identification appears more robust than others (this discussion could be added in section 4). (the dataset made available could therefore include two benchmark simulated series over the whole period, one using the parameter set calibrated on the first sub-period and one using the parameter set calibrated on the second sub-period.
We agree with this comment and think it is a great addition to the paper and dataset. The full Klemeš (1986) split sample calibration has been included in the dataset along with calibrations for two other forcing inputs (Maurer et al. 2002) and NLDAS-II (Xia et al. 2012) with some basic results in Figure 6 and discussion in the text. The various calibrated parameters and model output for the calibrations shown in Fig. 6 are also included in the dataset.

8. P. 5607, L. 4: As mentioned above, I do not understand the choice of this objective function, given its known limits (also acknowledge by the authors later in the text). Using a KGE-type objective function would also avoid useless discussions later in the article (section 4.2) on the limits of the proposed benchmark given the known problems of the selected objective function! Although I know other objective functions may be even more powerful, the advantage of KGE is that it remains very simple to compute. Note that I better understand however the selection of NSE as a criteria for model performance evaluation in this study to give this commonly-used performance reference.
See above discussion in response to major point D.

9. P. 5607, L. 18-22: It is useless to repeat in the text the information already given in the table.
We respectfully disagree on this point and feel it is worthwhile to discuss the calibrated parameters in the text along with the reasons why they were chosen.

10. P. 5607, L.22-25: This sentence is unclear.
We have tried to improve the clarity of this sentence.

11. P. 5608, L. 13-17: Indicate the units of each term of the equation.
We have included units of each term.

12. P. 5609, L. 1: What are these components?
We have noted the components.

13. P. 5609, L. 8-17: A similar climatological benchmark was advised long ago by Garrick et al. (1978) (see also Martinec and Rango, 1989). This could be mentioned. Why a 30-day smoothing window was deemed necessary compared to the simple reference proposed by Garrick et al. (1978) that simply uses the averaged measured discharge from past years for each day of the period? Is there any difference between these two references in terms of performance?
We thank the reviewer for this comment. It was very helpful looking at these two papers and we have included reference and discussion of them in the text. The 30-day smoothing window was used to provide a smoother daily long term climatology. Just using the 30 values on a given day may be highly influenced by one event, while the smoothed time series provides a historical benchmark more representative of a monthly climatology. We have not examined the difference, but there will likely be a small difference between the two approaches.

14. P. 5610, L. 4-14: This paragraph would probably be better placed at the end of section 2 with a more in-depth analysis of catchments characteristics (see comment above).
We have moved this paragraph and greatly expanded the discussion of the basin characteristics (see response to minor comment #1)

15. Section 4.2: Results on MNSE could be commented in the text.
We have included discussion of MNSE in the text.

16. P. 5611, L. 12-15: I do not agree with this argument. The fact that NSE is widely used does not justify that it should be used here, given it was demonstrated to be a bad choice for model calibration. I think this choice is even counterproductive for the community, since it will encourage a statu quo in the use of RMSE for model calibration if one wants to compare results with the proposed benchmark. I really think the use of KGE-type objective function should be encouraged. (note that I am not one of the developers of the KGE criterion, but I find it useful in practice).
We respectfully disagree with this assertion. Presenting results using RMSE as the objective function and using NSE as the main performance metric give the opportunity to use more advanced calibration methodologies

17. P. 5614, L. 2-6: This issue of data quality in climatic data may also be commented in section 2.2.
We have included mention of this in section 2.2

18. P. 5614, L. 23-25: Indicate if these are calibration or validation results.
We have clarified this statement

19. P. 5615, L. 4: What is a “low-order” hydrologic model?
We have clarified this statement

20. P. 5620, L. 29: Maybe not so useful to cite a paper in preparation if it is ultimately not published and therefore not possible to find it for readers.
We have removed mention of this paper.

21. Fig. 1.b: What RAIM stands for? An interesting graph would also be to plot the ratio of mean flow and mean precipitation (y=Q/P) as a function of the ratio of mean precipitation and mean potential evapotranspiration (x=P/PET). The graph could show the limit lines y=1 and y=1-1/x. The advantage of this graph is that it is based on observations only, whereas the graph shown by the authors uses model estimates.
RAIM stands for Rain plus Melt (RAIM). We have added explanation of the abbreviation in the figure caption. We have also added a panel of the proposed figure along with discussion in the text. Just to note, PET is not truly observations only as Daymet estimates shortwave down via regression equations found in MT-CLIM.