Hydrol. Earth Syst. Sci. Discuss., 11, C2280–C2285, 2014
www.hydrol-earth-syst-sci-discuss.net/11/C2280/2014/
© Author(s) 2014. This work is distributed under
the Creative Commons Attribute 3.0 License.

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Development of a large-sample watershed-scale hydrometeorological dataset for the contiguous USA: dataset characteristics and assessment of regional variability in hydrologic model performance" *by* A. J. Newman et al.

Anonymous Referee #2

Received and published: 8 July 2014

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:

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.

C. The presentation of the data set could be improved, by introducing a more detailed description of the catchment physical characteristics.

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 I think this makes the proposed benchmark a bit outdated.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.

I also give a number of minor suggestions below. Since I think new calculations would be useful to improve the manuscript, I advise major revision, but I fully support the ultimate publication of this work.

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.

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.

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?

4. P. 5603, L. 23: write “contiguous United States (CONUS)”

5. P. 5603, L. 12: What “MT-CLIM” stands for?

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?

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.

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.

9. P. 5607, L. 18-22: It is useless to repeat in the text the information already given in the table.

10. P. 5607, L.22-25: This sentence is unclear.

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

12. P. 5609, L. 1: What are these 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?

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).

15. Section 4.2: Results on MNSE could be commented 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).

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

18. P. 5614, L. 23-25: Indicate if these are calibration or validation results.

19. P. 5615, L. 4: What is a “low-order” hydrologic model?

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.

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.

Cited reference

Garrick, M., Cunnane, C., and Nash, J. E.: A criterion of efficiency for rainfall-runoff

models., J. Hydrol., 36(3-4), 375-381., 1978.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377(1-2), 80-91, 2009.

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., and Andréassian, V.: Large-sample hydrology: a need to balance depth with breadth, Hydrol. Earth Syst. Sci., 18(2), 463-477, 10.5194/hess-18-463-2014, 2014.

Klemeš, V.: Operational testing of hydrological simulation models, Hydrol. Sci. J., 31(1), 13-24, 1986.

Martinec, J., and Rango, A.: Merits of statistical criteria for the performance of hydrological models, Water Resources Bulletin, 25(2), 421-432, 1989.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 11, 5599, 2014.