Review of

“How many models do we need to simulate hydrologic processes across large geographical domains?”

By Wouter J. M. Knoben, Ashwin Raman, Gaby J. Gründemann, Mukesh Kumar, Alain Pietroniro, Chaopeng Shen, Yalan Song, Cyril Thébault, Katie van Werkhoven, Andrew W. Wood, and Martyn P. Clark

This study is a synthesis of (1) observed and simulated data from a study using the CAMELS dataset and a subset of models from the Modular Rainfall Runoff Modelling Toolbox (MARRMoT) [1], and (2) the `gumboot`-methodology for postprocessing the residuals errors models using a mixture of Bootstrap and Jacknife methods [2] of the calibration and validation periods based on NSE and KGE performance metrics. The postprocessing reveals a high variability of the sampling uncertainty among the models. This can be used as an additional criterion to assess the model quality, and it supports the selection process when large domains are modeled with a lower spatial resolution.

The results of this study are particularly significant, as the single use of integrated metrics such as NSE and KGE often leads to significant equifinality among potential models, which makes model selection difficult. The statistical method used to analyze differences in performance and sampling uncertainty may improve model selection and, thus, good modeling practice in the future. This study shows evidence of the applicability of the concept for large domains modeled with a lower spatial resolution.

The paper is within the scope and very interesting for the readers of HESS. The authors address a topic of high relevance, which significantly contributes to improving good modeling practice.

The authors have done a commendable job presenting the scientific results concisely and well-structured. I have only minor issues which should be addressed before publication:

INTRODUCTION:

- I see “Bayesian model averaging and selection” as a paradigm of equal importance as the “single model approach” and the “multi-model mosaic approach”. The latter differs from the more rigorous “multi-model Bayesian paradigm” because it seems based more on professional expertise than statistics. So, the Bayesian paradigm should already be discussed in Section 1.1.
- The introduction mainly focuses on the challenges when only streamflow observations are considered output variables. This limitation should be highlighted here or in the LIMITATIONS-Section.

LINE 165:

- Reformat “gumboot”.

LINE 161:

- Please give the full configuration of the application of the `gumboot`-methodology, such as time period, block size, number of blocks, number of samples … Is the time period different from the one used in [2]?

LINE 171:

- Please give a formal definition of the linear program solved here.

LINE 181:

- I suggest moving the following lines to RESULTS-Section.

FIGURE 2d:

- Could you highlight the best “performance-equivalent” models in red?

I suggest that the authors consider the above points before final publication. This will ultimately benefit the manuscript and the overall study.

[1] KNOBEN, W. J. M.; FREER, J. E.; PEEL, M. C.; FOWLER, K. J. A. & WOODS, R. A.: A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments. In: Water Resources Research 56 (2020), Nr. 9

[2] CLARK, M. P.; VOGEL, R. M.; LAMONTAGNE, J. R.; MIZUKAMI, N.; KNOBEN, W. J. M.; TANG, G.; GHARARI, S.; FREER, J. E.; WHITFIELD, P. H.; SHOOK, K. R. & PAPALEXIOU, S. M.: The Abuse of Popular Performance Metrics in Hydrologic Modeling. In: Water Resources Research 57 (2021), Nr. 9