Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/hess-2021-58-RC1, 2021

Comment on hess-2021-58

Anonymous Referee #1

Referee comment on "Technical note: PMR – a proxy metric to assess hydrological model robustness in a changing climate" by Paul Royer-Gaspard et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-58-RC1, 2021

I have completed my review of the technical note "PMR - a proxy metric to assess hydrological model robustness in a changing climate", by P. Royer-Gaspard et al., submitted to HESS. The paper presents a metric for evaluating model robustness in estimating flow volumes over disparate historical climate conditions. It then applies this metric to a set of 337 catchment models and compares it to the conventional differential split-sample test (DSST) approach.

The paper is concise, clear, well-written and well-organized. The readily calculated metric has the potential to be of use in model intercomparison studies, for characterizing model robustness in model selection, and potentially for multi-objective calibration. I recommend acceptance subject to moderate revision, as outlined below. I have also included minor comments in the supplemental document as pdf.

1) My one main issue with the evaluation approach used here is in the exclusive use of the absolute model bias from the DDST as an 'default' indicator of robustness, with the expectation that if the PMR metric is correlated to the absolute model bias (determined from DSST testing), then the PMR is an adequate proxy for robustness. The problem with this is in the use of absolute model bias. I will here address this via an example. In a standard DSST, the model is calibrated to a period of the historical record and validated to another period. Performance is deemed "robust" if the performance is minimally sensitive to the characteristics of the calibration and validation periods. For instance, if a model calibrated during wet years and validated during dry years exhibits similar validation performance (in terms of NSE, KGE, Bias, etc.) than the same model calibrated during dry years and validated during wet years, then it would be deemed robust to changes in climate. Thus, if these two model configurations both had a percent bias of 20%, the model is robust to changes in climate, even if not particularly accurate. If one model configuration had a percent bias of 20% in the validation period and one of -20%, then the model is not robust – it exhibits strong sensitivity to climate conditions. However, this is not sensitivity that would be picked up in a comparison of absolute model bias as calculated using equation 2 nor is this sensitivity fully picked up by the raw value of model bias in validation, which is a measure of accuracy rather than robustness (though I

recognize that a robust model should ideally minimize the variance of this model bias on an annual basis). A better indicator of robustness in this context might be the absolute difference in bias exhibited by the two alternate configurations of the model, e.g.,

$$\left|\frac{\bar{Q}_{sim,i}}{\bar{Q}_{obs,i}
-\frac{\bar{Q}_{sim,j}}{\bar{Q}_{obs,j}\right|$$

where i and j denote the dry/humid or warm/cold sub-periods periods. While I am not averse to the additional comparisons made to the absolute bias metrics, these are not themselves particularly strong indicators of robustness because they don't compare two different climate conditions – the whole value of the DFFT. I think that the authors need to therefore use a more appropriate DFFT-derived robustness metric (such as this one) as an additional basis for comparison. Because they have already done the analysis herein and would only have to post-process model results, I hope that such an addition would be relatively straightforward, and could add much to the paper.

2) I also believe that the authors should make it clear that this metric only addresses one form of model robustness – robustness in estimating annual volumes. Other approaches would be needed to examine robustness with respect to peak flows, baseflows, etc.

3) Lastly, this analysis should really have been carried out in terms of water years rather than Julian years, but I see no reference to this in the text. It would be appreciated if this could be clarified.

Please also note the supplement to this comment:
https://hess.copernicus.org/preprints/hess-2021-58/hess-2021-58-RC1-supplement.pdf