Reply to Anonymous Referee 1

Paul Royer-Gaspard, Vazken Andréassian, Guillaume Thirel

We would like to thank the reviewer for having accepted to review our technical note and for the constructive comments on the manuscript. We answer below the reviewer's comments and we also attach the annotated manuscript with answer to the reviewer's specific remarks.

Comment

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.

Reply

By building a composite DSST experiment, we wanted to diversify the reference and test various transfer conditions. We like your idea, which consists in implementing a metric that directly focuses on pre-defined transfer conditions in a straightforward way. Such a metric might be well suited to assess more accurately which types of climatic change a hydrological model struggles with. We will test it in the revised version. However, we would like to stress that our prime intention is to suggest a synthetic proxy for model robustness, which does require a minimum amount of hypotheses, and especially which does not require more than one model configuration, as is the case with physically-based models.

Comment

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.

Reply

The reviewer is right: in this note, we only address model errors on annual volume, which are usually a major concern in many studies about model robustness. We have formulated the PMR based on the work of Coron et al. (2012, 2014) who focused on model bias. However, our idea is that the PMR could be further adapted to other modelling exercises. We see a few key aspects to

compute the PMR: i) the possibility for model bias on a given sub-period to be either higher or lower than average model bias, ii) the possibility to interpret easily its value (which would not be the case if its computation was based on NSE for example), iii) the fact that it gives a rough idea of the bias obtained in a split-sample test. Any modification of the PMR that would respect these requirements would be suitable. Actually, in the PhD of the first author (soon published though written in French), the PMR has been adapted to various flow ranges in order to give more insights in model weaknesses in high and low flows. One could also imagine applying transformations to streamflow values (such as power, inverse, etc.) to decrease/enhance the weight of some hydrological events. This technique is quite popular in model calibration applied with KGE or NSE. For this technical note, we thought that adding these aspects would add too much complexity, but we will add a small paragraph in the discussion section.

Comment

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.

Reply

The reviewer is entirely right. We did carry out the analysis in terms of water years (for France October to September). We will make it clear in the revised version.

We hope that our answer will help clarifying unclear points, and want to thank you again for initiating this discussion.