Reply to Anonymous Referee 2

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.

Comment

First, while the motivation for such an indicator is framed in the context of climate change impact assessment and the simulation of conditions different from those observed, the metric is limited to assessing transferability within the context of observed climate variability/change. In the conclusion it is stated that the new metric can be used to help select models for climate change impact assessment. I think it should be clarified that just because a model is more robust in the period of observations, that does not necessarily mean it will be robust to changes outside the range of variability in those observations.

Reply

We definitely agree with the reviewer about the distinction between past variability and future climate conditions. Model evaluation based on our metric, as well as on split-sample methods such as the DSST, is a necessary yet not sufficient requirement for a successful use of a model in climate change impact assessment. We will stress this point in the revised version of the technical note.

Comment

Upon reading I was left wondering how the idea of moving biases might help inform model selection. The authors indicated that they saw little clustering of biases by indicators of catchment topography or climate. Did this assessment (not presented) include information on groundwater storage which may be more challenging for GR4J to capture.

Reply

No strong evidence in catchments properties was found to explain potential mismatches between PMR values and bias values obtained in DSST. Among the properties tested, there were topographical indexes such as catchment surface, mean and max elevation, mean steepness, or forest cover, as well as hydro-climatic characteristics computed from daily streamflow, temperature and precipitation data. The possible influence of groundwater on the hydrological regime has been solely evaluated through the use of a baseflow index. No other sources of data have been used. However, we understand that the reviewer question seems more related to the link between lack of robustness as assessed by the PMR and catchment properties, rather than the link between PMR values and biases obtained in DSST. We have not included such an analysis in our paper since our intention was not to evaluate GR4J but rather to present a new metric. We will though consider adding a paragraph on this issue in the manuscript.

Comment

The authors identify sub periods without any recognition of drivers of climate variability and their periodicities. Might it be possible to condition selection of L based on predominant modes of variability, eg. NAO?

Reply

We like the reviewer's idea of choosing sub-periods length with focus on the variability of large-scale climatic drivers such as the NAO index (or other relevant indices for the region in focus). This would help selecting models which may correctly simulate hydrological response to shifts in climate forcings for climate change impact assessments. Although sub-periods length in split-sample testing studies is usually chosen according to data availability (e.g. to divide a time series into 5 sub-series) or to calibration requirements, some authors have informed this choice with respect to hydrological events which models may have difficulties to simulate (e.g. sustained droughts, Fowler et al., 2016). While hydrological events have the potential to stress model failures and are essential for model development, our intention with the PMR is to suggest a metric which may offer a more general insight of model robustness by evaluating models ability to simulate the effect of climate variability on the hydrological regime. The reviewer's suggestion suits this objective well. Although we will not include such an analysis in the technical note, we will write a few lines about this topic in the

discussion section.

Comment

I agree with the other reviewer on the limitation of examining absolute bias and urge the authors to consider the solution offered. I was also a little concerned that using the average to limit the effect of 'drastically wrong' years may in fact be ignoring the most informative data points we have. Understanding why such years are so poor surely offers important insight. Perhaps some further discussion of this could be offered.

Reply

We will test the metric suggested by the first reviewer. As discussed in Appendix B, other mathematical forms than the absolute average could be used to increase the influence of 'drastically wrong' years. Since our idea is to suggest a synthetic metric about model robustness on the whole observed past variability, we would still advise using the absolute average. Modellers willing to get further insights in model robustness could compute both the PMR as we suggest it (to get the wider picture) and the metric suggested by the first reviewer (to assess model robustness in particular hydro-climatic changes).

Comment

Does the metric assume that the observational uncertanties are stationary?

Reply

We have made no hypothesis on measurement errors, since we focused on the way model errors should be accounted. Thus, indeed, measurement errors are implicitly considered stationary in time. Please note that this hypothesis is common to most of DSST studies.

Comment

Finally, the aim of the modelling exercise is often important to study design. The metric is limited here to assessing annual flows. How might this be used if the emphasis of the modelling study is on low flows, or high flows under climate change for example.

Reply

Concerning this last comment, the reviewer is kindly invited to read our answer to the second comment of the first reviewer.

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