Technical note: PMR – a proxy metric to assess hydrological model robustness in a changing climate

Paul Royer-Gaspard, Vazken Andréassian, Guillaume Thirel

Answers to the comments of the reviewers

We answer here formally to the remarks of the reviewers.

We answer straightly to the questions (the thanks were already expressed during the discussion).

The color code is the following:

- the review is in black ;

- our answers are in blue ;

- the modifications introduced in the paper are in red.

Reply to Anonymous Referee 1

1 Major comments

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...

Reply

We thought that your way of presenting model robustness was worth adding in the manuscript. We used it to improve our explanation of the PMR and what it represents. We have also changed some of the paragraph describing the whereabouts of its computation.

L.72-81 have been changed:

~~A robust model should thus adequately simulate streamflow volumes for any type of climatic conditions experienced by a catchment~~

Performance is deemed “robust” if it is minimally sensitive to the characteristics of the calibration and evaluation periods. For instance, if a model calibrated during wet years and validated during dry years exhibits similar validation bias than the same model calibrated during dry years and validated during wet years, then it would be deemed robust to changes in climate. A robust model should thus simulate streamflow volumes for any type of climatic conditions experienced by a catchment with a stable bias (of course, the lower the bias, the better). For example, 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. It should be noted that a model may lack of robustness while providing accurate (i.e. unbiased) estimation of average streamflow volumes on a long period of time.

We added on L.114-121:

As explained in the paragraph 2.1, the idea behind the PMR is that the robustness of the model is linked to the variability of model performance in time. By computing the difference between the moving bias curve and model average bias, the metric track changes in model bias across time around its mean value. It should be noted that, if the evaluated model is unbiased, as is the case on Figure 2 for the gray moving bias curve (model calibrated on the whole data), then the PMR reduces to the absolute integral of the moving bias curve around “$y = 0$.” Although the terms in the sum are computed in absolute terms, this does not mean that changes of sign in model bias (for example from 20% to -20%) are not accounted. Indeed, we should recall here that the PMR computes the deviations of model bias from model average bias: thus, any variations of model bias contributes to the PMR, whether actual biases on the sub-periods are negative or positive.

Some changes were also provided on L.122-127:

~~The reason for including a factor 2 is to reproduce the bias that would be obtained in a DSST on sub-periods that are on the opposite side of the moving bias curve (see Figure 2). Although the errors for each sub-period are calculated in absolute terms, the normalization by the average observed streamflow allows the resulting value to represent the average relative error produced by the model on the sub-periods as compared with mean observed streamflow.~~

~~One reason for computing errors relative to the average streamflow on the whole time series instead of the average streamflow of each sub-period is that this reduces the weight of very dry years. It also avoids dealing with zeros in the denominator in intermittent catchments. This choice is further~~

~~discussed in Appendix B.~~

In order to compare the PMR with model biases in DSST, we included a factor 2 in the computation of the PMR, in order to compensate the smoothing effect of comparing model biases on sub-periods to the average model bias (see for example the gaps between the red and blue moving bias curves on Figure 2, compared to accounted deviations from the gray moving bias curve). Normalizing by the average observed streamflow instead of the average streamflow of each sub-period was decided in order to reduce the weight of very dry years. It also avoids dealing with zeros in the denominator in intermittent catchments. This choice is further discussed in Appendix B.

In addition, we would like to discuss further your following sentence: *"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."*

You may have noted that we have avoided mentioning NSE and KGE in the added paragraph (L.72-81). Indeed, we think that the values of such squared-error-based metrics are highly sensitive to the type of river regime in consideration. A model would usually show lower NSE or KGE scores on periods with high flows than on periods with lower flows because of model error heteroscedasticity. In this context, comparing NSEs (or KGEs) obtained in validation on two distinct periods (NSE on period A after calibrating on period B, NSE on B after calibrating on A) may lead to inappropriate interpretation of model robustness. It is indeed possible that a low NSE score in validation on period B is actually very close to the highest possible score on this period, and that a high score in validation on period A is also very close to the highest possible score on period A. In this case, NSEs in validation on A and B would be different but we would consider that the model is robust (although inaccurate on period B).

This argument could be applied on bias as well, since it also depends on the average streamflow conditions. However, this dependency is straightforward (which is not the case for NSE or KGE). Besides, models are usually unbiased in calibration, because usual objective functions explicitly target bias and because most of the models have free parameters able to correct their simulated water balance. It is therefore easier to compare model biases on different periods because they share a common reference (i.e. bias=0).

Although we consider that this discussion is not worth adding in the manuscript, we have written a short sentence in the discussion section about this issue (L. 305-307).

Comment (following the previous one)

... 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

We agree with your point. In the case of GR4J calibrated with KGE, bias in calibration usually are close to zero and we had thus neglected this issue. For the sake of generality, we have changed our default robustness according to your suggestion (L.175-186).

~~It should also be mentioned that model biases obtained in DSST have been accounted as the absolute differences to 1 so that they could be compared to PMR values, as follows:~~

$$\text{~~Model bias on subperiod } i \text{ (in \%)~~} = \left| 1 - \frac{\overline{Q}_{sim,i}}{\overline{Q}_{obs,i}} \right|$$

~~A drawback of this way of computation is that it removes the sign of model errors. The sign of model errors in the different DSST setups has therefore been analysed in Appendix A. In the following, model bias obtained in DSST will systematically be accounted for in absolute terms without further notice.~~

It should also be mentioned that model biases obtained in DSST were calculated with respect to model bias in calibration so that they address the stability of bias, and thus could be compared to PMR values, as follows:

$$\text{Absolute Model Bias on subperiod } b \text{ (in \%)} = \left| \frac{\overline{Q}_{sim,b} - \overline{Q}_{obs,b}}{\overline{Q_{obs}}} - \frac{\overline{Q}_{sim,a} - \overline{Q}_{obs,a}}{\overline{Q_{obs}}} \right|$$

The index a indicates the calibration period (i.e. the dry period when validating on the humid period, etc.). Please note that in the case of GR4J, biases in calibration are usually very close to zero because the model is calibrated by optimizing the KGE, which explicitly target model bias, and because GR4J has the ability to correct water balance with the free parameters governing intercatchment groundwater exchange. Therefore, the term on the right of the soustraction sign is negligible in practice. It should also be noted that since the PMR is positive by definition, model biases were computed in absolute values. A straightforward drawback is that it prevents interpreting the sign of model errors. Therefore, it has been analyzed in the different DSST setups in Appendix A. In the following, model bias obtained in DSST will systematically be calculated in absolute terms unless clearly stated.

Please note that bias computations were slightly modified in order to normalize model volumetric errors by the average streamflow on the whole period instead of the considered sub-period. This modification was made in order to compute model bias consistently with the way model error is accounted in the PMR. This change had little influence on bias values and on the equivalency between the PMR and absolute model bias in DSST. See for example Pearson's correlation from Table 2 (crossed out values correspond to those obtained with the former formulation).

DSST setup	'dry'	'humid'	'warm'	'cold'	'unprod'	'prod'	average bias
Correlation	~~0.48~~ 0.47	~~0.44~~ 0.44	~~0.53~~ 0.58	~~0.57~~ 0.52	~~0.71~~ 0.72	~~0.70~~ 0.65	~~0.76~~ 0.76

Your comment also warned us about the quality of the explanation of the PMR, because we were not sure whether you were concerned by the absolute bias in Equation 1, which would not allow to track a change of sign in model bias from a sub-period to another. We hope that the paragraphs we previously added to better explain the PMR approach will be enough so that you agree that the PMR focuses on the stability of model bias, whether actual model biases change sign in time, or not.

In addition, we have added a suggestion for an alternative to the PMR which would inform about i) the sign of model volumetric error and ii) inform about model robustness in specific transfer conditions. This alternative metric is presented in the discussion section (L.273-297, sub-section 5.2, "Predicting model bias in DSST from the moving bias curve", not copied here for the sake of brevity).

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

We fully agree with your point. We have added a small paragraph at the end of our introduction (L.58-62, copied below) as well as a dedicated paragraph in the discussion section about this issue (L.298-318, sub-section 5.3, "Generalization of the PMR for the evaluation of hydrological models", not copied here for the sake of brevity).

It is worth noting that hydrological model robustness is here considered especially through the prism of model bias. Given that the biased simulations are one the most common outcome of the previous works about model robustness, we considered that model bias was an adequate metric as a first approach. Of course, model robustness relates to the stability of model performance in general, and thus to every possible metrics assessing model skills. Hence, the PMR as presented here should be considered as *a satisfactory proxy* for model robustness as estimated using the DSST rather that *the proxy* for model robustness.

Finally, to compensate for the new discussion topics (sub-sections 5.2 and 5.3) and keep the technical note short and easy to read, we have removed the sensitivity analyses from the former sub-section 5.2 (Effect of sub-period length on the reliability of the PMR) and replaced it by the former Appendix C. These analyses were moved from the discussion section to the result section (i.e. in the new sub-section 4.3: "Influence of sub-period length on the sensitivity and on the reliability of the PMR").

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

You are right. We did carry out the analysis in terms of water years. We have made it clear in the revised version (L.144).

In France, a hydrological year lasts from October to next September.

2 Minor comments

L.101: “flatness of the moving bias curve” *Don’t you simultaneously need it to be flat and approach zero?? Couldn’t you indirectly shift the mean bias by forcing flatness?*

This would indeed be a possibility. We thought that it would be less confusing for model diagnosis to distinguish the curve’s flatness and model bias on the whole period, rather than mixing both information in a single performance score.

L.102-103: “We thus propose a simple mathematical expression to calculate this flatness, i.e., a performance metric designed to be a PMR.” *the PMR? You are going back and forth between calling this A PMR and THE PMR — you should pick one. Is this to be a class of metrics or a specific metric. Perhaps you could call this bPMR, a bias-derivived proxy for model robustness, opening the door for other future PMRs.*

You are right. The sentence was rewritten as follows:

→ We thus propose a simple mathematical expression to calculate this flatness, which will be referred to as the PMR in the following. Please note that the PMR is bias-derived and could thus be named PMR$_{bias}$, opening the door to other types of PMR based on alternative metrics. This issue will be further discussed in paragraph 5.3.

L.106-107: *While technically valid, the term ’flat curve’ always throws me off. Can’t you just say it is the difference between the moving bias curve and the mean bias?*

You are right. The whole paragraph was rewritten as follows:

~~The PMR is based on the computation of the average absolute difference between the actual moving bias curve computed on 5-year sub-periods and a hypothetical flat curve. This hypothetical flat curve is defined as the curve obtained for a hypothetical model, perfectly robust so that its bias on different time sub-periods would remain constant, but imperfect so that this bias would be equal to the mean bias of the evaluated model on the total period. It should be noted that, if the evaluated~~

~~model is unbiased, as is the case on Figure 2 for the moving bias curve obtained by calibrating the model on all the available data, then this reference hypothetical curve is simply defined by "y=0." The PMR is computed as the mean of absolute differences between the model average error on the 5-year sub-periods and the model average error on the total period, normalized by the average observed streamflow (Equation (1)). It thus corresponds to the normalized area between the moving bias curve and the hypothetical flat curve.~~

The PMR is based on the computation of the average absolute difference between the actual moving bias curve computed on k-year-long sub-periods and the average bias of the model, normalized by the average observed streamflow (Equation (1)). It thus corresponds to the normalized area between the moving bias curve and the average bias of the model.

L.112-113: "N is 100 the number of sub-periods that can be defined with a 5-year moving window" *Call this k, as it is potentially variable, and you can calculate N as M-k+1, where M is the total number of years of record.*

You are right. The sentence was rewritten as follows (see also L.107):

→ N is the number of sub-periods that can be defined with a k-year moving window (N=N$_{years}$-k+1 when there are no gaps in the data, with N$_{years}$ the number of years in the record).

See also (L.128-129):

→ In the following, sub-period length has been set to k=5 years. The choice of the sub-period length in the computation of the PMR is discussed in paragraph 4.3.

L.133: "(Delaigue et al., 2020)" *Explain context of citation; I presume this is a previous study for which these models were built/applied.*

You are right. The sentence was rewritten as follows:

→ The observed hydro-climatic data for the set of 377 French catchments used in this study (Figure 3) come from the hydro-SAFRAN daily dataset (Delaigue et al., 2020).

L.209: "model robustness" *Here and elsewhere, you are exclusively referring to robustness to volume error. I'm not comfortable with the suggestion that this is equivalent to general model robustness. Also this should be "a satisfactory proxy for model robustness as estimated using the DSST."*

This issue was addressed in the replies to the reviewer's major comments.

L.230: *Just an idea, and not required effort: It seems you could potentially apply a synthetic ex-*

periment where you introduce systematic non-robustness to artificial hydrographs and then test the PMRs ability to detect it.

This is an interesting idea. There is no perfect performance metric. Understanding the behavior of a metric is an important issue to properly interpret the results of an experiment.

L.322-323: "it is never used on models that are so complex that they need to be calibrated on all the available data." *They cannot be used on models that require calibration to the entire period of record. (this is not necessarily an issue of complexity, but could also be a limitation of data availability)*

You are right. This sentence now reads:

→ they cannot be used on models that need to be calibrated on all the available data or to uncalibrated models

L.341-351: *Does PMR have enough fidelity to deem an acceptable threshold? i.e, what is a sufficient/insufficient PMR value for models to be used in climate change studies? Some guidance here might be useful.*

In our opinion, defining an acceptable threshold for PMR is as difficult as it is for any other performance metric (KGE, RMSE, etc), because it depends on the catchment properties and on the modelling objectives. However, it may be interesting to adapt the metric in a skill score by comparing the performance of a studied model with the performance of a benchmark model. This is the idea behind NSE. Yet care must be taken to choose an appropriate benchmark. The mean model could be a possibility, as inspired from NSE. We would also suggest a yearly Budyko model. The whole paragraph was rewritten as follows:

→ Our results should encourage hydrological modelers to include the PMR as part of their panoply of evaluation metrics to judge their models. The metric addressing models transferability within the context of observed climate variability, it can be useful in model robustness assessments. In the context of climate change impact assessments though, it should be recalled that demonstrating model robustness in the historical period is a necessary yet not sufficient requirement to validate model robustness in future conditions outside the range of past observations. Still, being relevant for any kind of hydrological model, it may be used to inform model selection for such simulations. Of course, it appears difficult to define acceptability thresholds for the PMR a model should pass to be used in extrapolation, since it would be catchment and objective dependent. However, one could imagine adapting a standardized PMR by comparing PMR values with a benchmark model as is done for NSE (for example a simple yearly Budyko model). Further work should also examine the potential of PMR to be incorporated as a hydrological signature in multi-objective calibration procedures essentially to constrain model parameters governing slow temporal changes in catchment response.

3 Technical comments

L.6: "~~which performs with~~"

→ can be performed with

L.10: "~~adequation~~"

→ equivalency

L.17: "~~hydrological models~~"

→ hydrological model uncertainty

L.28: "~~couple~~"

→ pair

L.40: "~~were not~~"

→ cannot be

L.43-44: "~~define~~"

→ inform the definition of

L.56: "computation choices" More specificity welcome — add (e.g.,...)

→ We added (e.g. sub-period length, sub-period weight)

L.63: "~~obtained in~~"

→ as determined by

L.84: "model" *Model volumetric errors (this does not address timing or peak errors)*

→ model volumetric errors

L.195: “~~theoretically~~”

Done

L.223: “~~poorly~~”

→ less

L.326-327: “~~on a single model realization~~”

→ from a single model calibrated on the entire period of record

Figure 1: *Should align axis labels and axes in top and bottom figures. Since this is a moving window analysis, the centre of the window in the top plot should align with the corresponding point in the bottom plot.*

Done

Figure 1: *Is this using water years? It should.*

This issue was addressed in the replies to the reviewer’s major comments.

Figure 5: *Can we include these other scatterplots as smaller subfigures of fig 5? Just the Pearsons coefficient isn’t particularly informative.*

Done

Figure 5: *Average absolute bias.*

Done

Reply to Anonymous Referee 2

1 Major 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 fully agree regarding your concern about the distinction between past variability and future climate conditions. We added on L.341-345:

Our results should encourage hydrological modelers to include the PMR as part of their panoply of evaluation metrics to judge their models. The metric addressing models transferability within the context of observed climate variability, it can be useful in model robustness assessments. In the context of climate change impact assessments though, it should be recalled that demonstrating model robustness in the historical period is a necessary yet not sufficient requirement to validate model robustness in future conditions outside the range of past observations.

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 baseflow index. No other sources of data have been used simply because we conducted our analysis on a large sample of catchments about which we have no consistent set of available data, apart from the ones mentioned previously. However, I understand that your 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.

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 your idea of choosing sub-periods length with focus on the variability of large-scale climatic drivers such as the NAO index (or other relevant index 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. Your suggestion suits this objective well and we will have written a few lines about this topic in the discussion section (L.271-272).

One could imagine that it may be chosen according to the temporal variability or periodicity of some climate indices (e.g. the North Atlantic Oscillation index).

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 have replaced our former DSST metric by 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.

However, we fully agree with your concern about suggesting an alternative to the PMR in order to get further insights in model robustness in more specific transfer conditions. Thus, we have suggested a new metric, called sPMR (for *specific* PMR), in sub-section 5.2 (L.273-297), which is an adaptation of the metric suggested by the first reviewer.

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

We fully agree with your argument. We have added a few paragraphs about this issue (L.58-62 and L. 298-318). You are kindly invited to read our answer to the comments of the first reviewer for further details.