Transferring model uncertainty estimates from gauged to ungauged catchments

Dear Editor,

We would like to thank you and the five reviewers for the overall positive feedbacks on the article. We understood from the comments that several parts of the manuscript needed rewriting, clarification and further analyses.

As explained in the detailed response to the review comments, we agree with most comments made by the five reviewers. The main modifications we made in the manuscript to answer the major comments consist in:

- improving the description of the proposed approach, with the introduction of a general diagram of the method that intends to clarify the various steps;
- better explaining the evaluation method and criteria;
- providing a complementary analysis on the behaviour of the method in the case of data-scarce region, with an appropriate sensitivity analysis consisting in creating a “hydrometrical desert” in the gauging network around the target station;
- better discussing results and outcomes of the study and its possible implications, and also the possible limitations of the proposed approach;
- using a more consistent terminology throughout the paper and introducing definition of terms when necessary.

The revised version of the manuscript also account for all the specific changes requested by the reviewers, as explained below.

In the following pages, we shortly explain how we effectively accounted for each reviewer’s suggestion, following our initial reply (see bullet points ➢ in red).

D.A. HUGHES (R1)

While this paper presents an interesting and novel approach to quantifying uncertainty in hydrological modelling, I think that more should be said about the limitations of the approach. It has been applied in France, but I suspect that it would be difficult to apply in data scarce regions. There are many areas where there are simply not enough gauged catchments to represent the variability in the hydrological response across many ungauged catchments. A further problem is that many gauged catchments are also affected by poorly quantified anthropogenic impacts that will impact on the ability of the data to adequately represent the natural hydrological response that the model is trying to simulate. There are also potential problems with the lack of representativeness of the climate inputs in the gauged catchments that could lead to bias in the quantified parameter values of the donor catchments. These additional uncertainty issues do not seem to be addressed in the paper and I think that they should at least be mentioned and there impacts on the overall likely success of the method should be noted.

We thank Pr. Denis Hughes (reviewer R1), for his positive comments about the paper. We will take his comments into consideration to enhance the revised manuscript.

We agree that the approach is difficult to apply in data scarce regions and that anthropogenic impacts can lead to misleading results. But the same limitations apply to any regionalisation approach. We intend to include a complementary analysis in the discussion section, in which we will progressively decrease the density of donor catchments, to show the impact on uncertainty. This will help better discussing the applicability of the method in data scarce region. We will also discuss the issue of representativeness of the input and impact of human influences.

- We have included a complementary analysis with tests of a "hydrometrical desert" around the target station. This test is more demanding that decreasing the density of the donor catchments and reveals the limitation of the method.

Spatial proximity is mentioned on page 8045, but what about the effects of highly variable topography (or other factors) between closely adjacent catchments? Would this not invalidate an approach based solely on distance?

The selected approach for transfer of information from gauged to ungauged catchments is based on spatial proximity. This choice is motivated by previous work on regionalisation based on this data set. We are aware that physical similarity may be better adapted in some cases, as shown by other comparative studies on regionalization methods. However the proposed approach of uncertainty quantification could also be applied with any regionalisation strategy, for example physical proximity if it is deemed to be more appropriate. This will be further underlined in the discussion section.

- See the discussion section.

I did manage eventually to understand all of the steps in the method and the performance measures. However, I had to read them several times and I really think that they could be better explained. The paper is quite concise (generally a good thing), but in terms of the explanations I think it is too concise and would benefit from further and clearer explanations of some of the points within section 3 and 4. I refer to some examples below.

We agree that further and clearer explanations are needed to fully describe the method. We will modify the manuscript to make it more easily understandable and improve the graphical illustrations to support the explanation.

- See in particular the modification of the method section.

Page 8047 explains the sharpness index that used the Q5/Q95 ratio for the historical FDC. I think that the authors did not use the 'width', which would be Q5-Q95, and therefore should not refer to width. I also fail to see how 1-Q5/Q95 can provide a measure of uncertainty when it is solely based on historical flows according to the explanation provided in the text.

This is indeed not detailed in the manuscript. We will better explain how the sharpness index is calculated and how it is related to the mean width of the intervals, based on appropriate references.

- Done

Page 8048 refers to the skill or interval skill score. I think that the use of the 1{X,Y} notation is confusing in equation 1. Why not give this a variable name (e.g. INDF) and then use separate equations to define how INDF is calculated. I tried to understand what the skill score is doing and it seems as if high values of S relates to poor skill - is that correct, or did I get it wrong? I assume that l and u represent the lower and upper bounds of the uncertainty at any point in the time series? What is 'unconditional climatology'?

We also agree that further explanation is needed here. In particular, the equation will be rewritten in a simpler form.

- Done

Page 8050 refers to using donor catchments as gauged (the difference between this and treating them as ungauged also needs further explanation I think). Why should the results be less reliable in this case and why is there a benefit when treating donor catchments as ungauged - this seems to be somewhat counter intuitive?

As A. Viglione (R2) puts it, treating donor catchment as gauged is simply "wrong" because we have to expect that errors are larger in a regionalisation context as the errors obtained with calibration. As a consequence, the uncertainty estimates are less reliable because uncertainty is underestimated. But we agree that the two-step approach is not so intuitive and we will therefore improve the explanations on this aspect.

- See in particular the modification of the method section.

Page 8051: It was not immediately clear to me what data are used to calculate the NSE criterion? Is it the upper and lower prediction bounds or what? Please provide a clearer explanation.

We apologize for the confusion. We used the simulated discharge values to compute the NSE criterion as it is usually done. We will make this point clear in the revised manuscript.

- See in particular the modification of the method section.

I would therefore like to suggest to the authors that they seriously consider making the explanations for most of the methods a lot more clear so that readers can understand the approach and methods much easier.

This concern has also been expressed by most of the reviewers of this paper and we agree that we have to put more effort on the explanations. We will make appropriate changes to make the explanations of the methods a lot more clear.

- See in particular the modification of the method section.

Minor points and corrections:

P8042, L16: Surely this should be residual errors at gauged locations.

Our sentence was misleading and we will modify it. In the paper we cite, residual errors are first estimated at ungauged locations based on residual errors at gauged locations, and then quantile regression was applied with the estimated errors at ungauged locations.

- Done

P8043, L5: '.. of the work by Oudin...' P8044, L16: ' ... discharge data ARE available..'

Thanks.

- Done

P8044, L17: '.. ungauged LOCATIONS ...'

Thanks.

- Done

P8044, L25: Please indicate what the performance criterion is (NSE presumably)?

The performance criterion is the one used to calibrate the models, i.e. NSE computed on root square transformed flows. We will modify the sentence.

P8048, L24: Please use percentages (70%) instead of a fraction (0.7) to be consistent with the rest of the text.

Agreed.

- Done

P8049, L11: What is the basis for 30 and 80%? P8049, L13: 'yield' shoud be plural.

The values of 30 and 80% are arbitrary. We will add a sentence to make it clear.

- Done

P8049, L18: I do not understand what 92% represents nor where it comes from.

It was meant to be the fraction of catchments where ISS is positive. We will clarify this point.

- Done

P8050, L2: 'rainfall-runoff MODEL'

Thanks.

- Done

P8050, L26: 'increase' should be plural.

Thanks.

- Done

P8050, L26 & P8051, L3: should be 'compensate FOR..'

Thanks.

- Done

The lines used for the boxplots in figures 6 to 8 could be thicker to make them clearer in a printed version of the paper.

Thanks for noticing the issue. We will make the appropriate changes to get better figures.

- We include the figures in pdf format.

A. VIGLIONE (R2)

In this paper an estimation of the total uncertainty affecting runoff prediction in ungauged locations is performed. The total uncertainty is estimated based on residuals of the estimated runoff at neighbouring gauged catchments treated as ungauged (i.e., in cross-validation mode). I like the pragmatic procedure for the estimation of the total uncertainty. In fact, I was recently involved in editing a book on runoff prediction in ungauged basins (Blöschl et al., 2013, already cited in the paper), where, consistently with this paper, "total uncertainty" was assessed based on the performance of runoff prediction obtained in cross-validation over many locations (see also Parajka et al., 2013, already cited in the paper).

We thank A. Viglione (R2) for his positive comments about our paper and the pragmatic approach we presented. We will take his comments into consideration to enhance the revised manuscript.

One addition which, in my opinion, would make the results of this paper more interesting for the hydrologic community, would be to stratify the measures of reliability, sharpness and interval score as a function of climatic and catchment characteristics (i.e., aridity index, catchment area, catchment elevation, density of the gauging network, ...). In other words, is the method performing equally well in all France or are there problematic regions? If the latter is true, what could be the reasons? This could also serve to address the concerns of Reviewer #1 (Denis Hughes).

We thank R2 for his suggestion. We will carry out the suggested analyses and provide comments on the possible links. However, results from past regionalization studies in France showed it was very difficult to find regional trends or links between efficiency and catchment characteristics. This can be explained by the fact that modelling errors are manifold. So we are not sure convincing conclusions will be drawn on this aspect.

- We computed Spearman correlation between the C2M and the three probabilistic scores and P, Q, EP, P/Q and P/E, but did not obtain any value above 0.25. This indicates that it is difficult given those characteristics to provide further insights into the reason for the variability of performance we obtained.

The Authors have chosen to assess the reliability, sharpness and interval score for the 90% prediction intervals. Does the method perform equally well for other prediction intervals? More generally, since the method gives an estimation of the empirical distribution of the error for different flow groups, why haven't the Authors checked the goodness-of-fit of these distributions, for example through an uniformity test of the non-exceedence frequency of the actual error values (see e.g., Laio and Tamea, 2007, pages 1272-1273)?

We only presented the results obtained for the 90% prediction intervals; however the method can indeed be applied to obtain other prediction intervals and an approximation of the distribution. We believe that introducing the approach and

focusing on the 90% helps making the paper concise and easier to understand, even though most of the reviewers already pointed out that our presentation should be clearer. We agree that further work could be done in that direction. We will introduce corresponding comments in the concluding part of the article.

- Done

Overall I think that the paper is well written and sufficiently clear, even though I agree with Reviewer #1 in that some points could be better clarified. Even though I've asked to add some analyses, I think that a minor revision should be sufficient for that. Some specific comments follow.

Actually, the article will be quite deeply modified following all the comments received from the reviewers.

- See in particular the modification of the method section.

Page 8044, line 5: I would suggest to shortly discuss here in what are the two models different. I understand that this may be found in the previous papers, but for readability I would summarise the main differences here too.

Agreed. We will shortly discuss the main differences between the two models.

- A short description was added.

Page 8045, line 5: I have a concern about the "output averaging option". Since averaging many signals results into a smoother one (also in the case that they are correlated), are the extremes well predicted? If so, are the results in this paper affected by that? This could be checked, for instance, applying the procedure for the 98% interval.

The output averaging option concerns the regionalisation method used to obtain a deterministic prediction at the outlet of any catchment. As such it does not directly affect the procedure used to obtain uncertainty bounds. If for example the extremes are consistently underestimated at neighbouring locations, the procedure will be able to reflect such systematic bias. However, we agree that the choice of the regionalization option may affect the quality of simulation of some parts of the hydrograph. However, the proposed approach is not specific to a given regionalization setting and others could be adapted if deemed more appropriate. This will be clarified in the discussion section.

- Done.

Page 8047, Section 4: here the Authors introduce the concepts of "reliability", "sharpness" and "interval score". Regarding the first two, unless the concepts are new, which is not the case, I would suggest to add references here to where these concepts are extensively discussed (e.g., statistical books?).

Agreed. The concepts are not new and were used before in other publications. We will add references to previous work.

- Done

Page 8047, line 13: The sentence about the "two values" is a bit confusing here. The Authors intend the two average widths of the uncertainty bounds and of the historical flow quantiles, while at first I confused the two values to Q5 and Q95. I see that also Reviewer #1 had a problem with this sentence.

Agreed. The sentence was confusing and we will make the presentation of the sharpness index clearer.

- Done

Page 8047, line 15: What is the climatology?

By climatology we mean the unconditional distribution of observed values, i.e. the flow duration curve, from which we calculated the width of the 90% interval. That way, we obtain one value per catchment that reflects the natural variability. We agree that the presentation was unclear and we will make it clearer. A better definition of terms will be provided.

- Done

Page 8048, line 9: That's related to the previous point. I do not understand what a climatological interval is.

Agreed. We will make it clearer and we propose to add a new figure showing how it is calculated.

- Done

Page 8048, "interval score": I have difficulties to understand what S measures. Maybe some more information should be given to help the reader. I've seen that also Reviewer #1 has concerns about this.

The interval score accounts for both the width of an uncertainty bound and the position of the observed value compared to the uncertainty bound values. We will add a figure to show how the score is calculated.

- Done

Page 8049, line 21: Same here. What is the unconditional climatology?

Agreed. This point needs a clearer presentation.

- Done

Page 8049, Section 5.2: The results obtained using donor catchments as gauged are not surprising. They descend from the fact that the procedure is wrong, since calibration removes biases. It is interesting, though, to see the results from a wrong procedure. However I would stress in the section that the procedure would be "wrong" since the uncertainty of runoff prediction in ungauged catchments is of interest.

You are perfectly right, the procedure is "wrong" by construction because the magnitude of errors is not the same when calibration is used instead of the regionalisation procedure. We propose to add a new figure to clearly show how the performance of the two models decreases when we move from calibration to regionalisation.

- Done, see Figure 3.

ANONYMOUS REFEREE #3 (R3)

Overview:

The aim of the paper is very clear: estimate global uncertainty of the model output in ungauged catchments. Overall the paper is well-structured. It is also concise, which in general in a good thing. However, at certain points throughout the text further explanation would be helpful to aid interpretation.

We thank the anonymous reviewer R3 for his positive comments about the paper. We will take his comments into consideration to enhance the revised manuscript.

Main Points:

1) The Authors aim to estimate total uncertainty. However, in the text (including the title of the paper) they often refer to total/global uncertainty as 'model uncertainty'. The Reviewer thinks this can be misleading, as it sounds like the Authors are trying to assess the uncertainty introduced by the choice of the rainfall-runoff model.

The terminology used in the context of uncertainty estimation is indeed sometimes confusing, and R5 also pointed out this issue. We agree that the procedure we presented aim to estimate total/global uncertainty. We propose to modify the title of our paper and the expression "model uncertainty" by "global uncertainty".

- Done

2) The Authors suggest a way to estimate total uncertainty in an ungauged catchment based on neighbouring gauged catchments. Although the Reviewer does not have a problem with this, the way the Authors implemented this methodology may be faulty. Using the catchments shown in Figure 1 as an example, the errors estimated for the green catchment resulting from transferring information from the yellow catchments (figure 1 B) are probably not representative of the errors expected from the transference of the information from the red catchments to the grey catchment (Figure 1 A). The errors calculated for the green catchment based on the yellow catchments are likely to be smaller as the catchments seem to be nested. On the contrary, the prediction of the runoff hydrograph of the grey catchment uses four catchments from different river branches and therefore the Reviewer expects that the error in this case is higher. Therefore, the Reviewer believes that the way the catchments were selected to estimate the uncertainty is not adequate.

R3 rightly points out that we did not take into account the fact that some catchments are nested. This could be done within the framework of our methodology. We do not have any expectation regarding the fact that the errors are higher or not in these cases but we will mention that further work could be done to investigate this issue. Note that we agree that the example used to illustrate the approach may introduce some confusion on these aspects and we will therefore use an example without nested catchments.

- We changed the illustration of the method.

3) The paper lacks a critical evaluation of the methodology suggested.

We are not sure to understand what R3 means here. We believe that applying the methodology on a large set of catchments and using a quantitative evaluation with three widely used and recognized scores of the obtained uncertainty bounds is a way to rigorously evaluate the methodology. But as suggested by other reviewers, we will better discuss the possible limitations of the proposed approach.

- We acknowledge some limitations in the conclusion.

Minor points:

1) American English and British English are used interchangeably. Some examples (among many others) include: on page 8040, line 21, 'modelling'; on page 8041, line 16, 'behavioural'; on page 8044, line 10, 'optimization'; on page 8051, line 9, 'characterize'.

Thank you for pointing out this issue. We will make adequate modifications to correct the mistakes in our manuscript and use more consistent language writing.

- We tried to correct the mistakes but we acknowledge that our English is perfectible. Therefore, we will ask a professional to carefully check our wording if the paper reaches the publication phase.

2) Page 8040, lines 24-25: What do the Authors mean by 'prediction approaches'?

We apologize for the misunderstanding. By "Bayesian calibration and prediction approaches" we mean the application of Bayes theorem to infer unknown values and then propagate the uncertainty sources for prediction.

- Precision introduced

3) Page 8041, line 10: What are the parameter sets constrained on?

Parameter sets can be constrained by various sources of information, including the regionalized "signatures" and soft information, as mentioned lines 14-15.

- No modification necessary

4) Page 8041, line 14: hydrographs or hydrograph?

Thanks. We mean hydrograph.

- Done

5) Page 8041, line 16: How does the second step relate to the first step?

The first step provides regionalized metrics used in the second step where only some parameter sets - the ones that provide metrics close to the regionalized metrics- are retained. This will be clarified.

- Done

6) Page 8041, lines 10-19: In a Bayesian approach, like Bulygina et al. (2012) used, there is no distinction between 'acceptable'/'behavioural' and 'non-behavioural' parameter sets. All parameters are acceptable, though some are more likely than others. Therefore, the Reviewer suggests the Authors to rewrite this sentence.

Agreed. We will rewrite the sentence to introduce the distinction between formal and informal approaches.

- Done

7) Page 8042, lines 9-12: This is an example of where the Authors were too concise resulting in an explanation that is not satisfactory. Before reading the rest of the paper, and solely based on this paragraph, it seems that the Authors are suggesting that neighbouring gauged locations are calibrated and the residuals between model prediction and the observed data at these catchments are used/transposed to the ungauged catchment for uncertainty estimation at this location. The Reviewer does not agree with this, as in the ungauged problem there are additional sources of uncertainty when compared to the gauged problem. For instance, additional sources of uncertainty introduced by the transference of information should be taken into account when the final goal is to estimate the global uncertainty of the model output in the ungauged catchment. This is, in fact, highlighted later on by the Authors (Figure 7 and Section 5.2, page 8050, lines 1-3). This needs to be more clearly explained in the early stages (e.g.Introduction) of the paper.

Thanks for pointing out this issue. The sentence can indeed introduce some confusion and we will modify it in the revised paper.

- Done

8) Page 8042, line 21: are instead of is.

Thanks.

- Done

9) Page 8044, line 2: Why did the Authors select 4 and 7 catchments? What is the justification for using these particular number of catchments?

In this paper, we chose to adopt the options in the application of the regionalisation method, which were selected by Oudin et al. (2008) in their previous study on the same data set and the same models. We purposely considered that the regionalisation procedure is given and we focused on the uncertainty quantification issue only. However, we wanted to present an approach that could be used with any regionalisation strategy. This will be more clearly stated in the paper.

- Done

10) Page 8047, lines 9-21: In general, the definition of sharpness is confusing and should be clarified. The Reviewer interpreted AWI as being [1-average width uncertain bounds/(Q95-Q5)], but this should be better explained. In particular, it is not clear which 'two values' the Authors are referring to on line 13. It is also not clear what the Authors mean by 'compared to the climatology', in line 15. In line 16, what is the percentage reduction of the average width in relation to? Line 17, reduced in relation to what?

Agreed. Similar comments were made by other reviewers. We will modify the paragraph to better define the evaluation strategy.

- Done

11) Pages 8047-8048, Equation 1: It may be worth explaining what range of values would be expected for S, which values correspond to a poor prediction and which values correspond to a better prediction.

Agreed.

- Done

12) Page 8048, line 1: It may be worth clarifying what 'l' and 'u' are.

Agreed.

- Done

13) Page 8048, line 5: What does 'unconditional climatology' mean? Please clarify.

Agreed. By climatology we mean the unconditional distribution of observed values, i.e. the flow duration curve, from which we calculated the width of the 90% interval. This will be clarified in the manuscript.

- Done

14) Page 8048, Equation 2: The Authors have used ISS on the left and on the right hand side. The Reviewer assumes that on the right hand side it should be IS instead of ISS. Please correct this, if that is the case.

Thanks for pointing out this mistake in the equation. We will modify it.

- Done

15) Page 8048, line 11: Do the Authors mean skill score (IS) or interval skill score (ISS) here?

We mean the interval skill score (ISS).

- Done

16) Page 8048, lines 21-23: The Authors say that the median values for reliability for GR4J and TOPMO are 89% and 90% respectively (also shown in

Figure 6). Roughly half of the catchments are above the expected 90% value for the 90% prediction bounds, and the other half is below. Therefore, the Reviewer is of the opinion that the Authors should not say that "the prediction bounds are, in most of the cases, able to reflect the magnitude of the errors", when those cases represent only 50% of the cases. The Reviewer suggests that 'in most cases' should be changed.

Agreed. We will modify the sentence so that our presentation of results does not appear too optimistic.

- Done

17) Page 8048, line 24, and page 8049, lines 1-3: This comment links with comment 16. Why do the Authors use CR=0.7 as a benchmark, when they say beforehand that 0.9 should be expected for reliability? Using CR=0.7 as a benchmark is misleading as it makes the results seem better than they actually are. If the aim here is to estimate total uncertainty and a value of 90% is expected for 90%prediction bounds, the Authors should focus on CR=0.9. As said before, approximately half of the catchments present a CR<=0.9, indicating that for 50% of the cases the uncertainty bounds might be too narrow or biased.

Agreed. The choice of using a value of CR=0.7 is arbitrary, and a perfectly reliable methodology used to quantify uncertainty should yield a value of CR=0.9. In fact it is difficult to find in the literature any guidance about how to evaluate properly the CR values. We propose to add a few sentences to discuss this issue. We will also make clearer that the results show some limitations of the proposed approach. Nonetheless, we believe that the results shown in this study could be used by other teams as a general benchmark.

- Done

ANONYMOUS REFEREE #4 (R4)

This paper deals with the highly challenging and important problem of quantifying uncertainty in streamflow estimates at ungaged locations. I think this paper moves forward the discussion on this topic by providing a novel and practical approach and is, therefore, suitable for publication in Hydrology and Earth Systems Science. The manuscript is well-written and I have only minor editorial comments. I do also have some major comments/questions that could improve the clarity of the manuscript.

We thank reviewer R4 for his positive comments about the paper. We will take his comments into consideration to enhance the revised manuscript.

Major comments/questions:

1. Could the authors make some clarifying comments about the difference between confidence intervals/estimated and prediction intervals/estimates? It seems to me that that the early part of the experiment presented here focuses on the confidence intervals/estimates around estimated streamflows and the latter portion of the work (Section 5.2) as an attempt to define the prediction intervals of the estimated streamflows. Is this what the authors intended?

We use the term "prediction intervals" to describe intervals aiming at describing uncertainty around a deterministic value. In particular, prediction intervals are expected to cover the range of variability of the target variable, while confidence intervals do not. Note that there is no difference amongst the different experiments presented here. We will clarify this confusing point.

- We introduced a new section and statements to clarify the different experiments presented. We hope that it helps to make it clear that we only focus on prediction intervals. Thus, we did not feel that other modification were necessary to explain the difference between confidence intervals and prediction intervals.

2. If the authors were intending to obtain prediction intervals for the estimated streamflows, then only the experiment design for Section 5.2 seems valid to analyze here. More clarifying statements are needed to understand why the experiments were done both ways (treat donors as gauged or ungauged).

We did the two experiments because we wanted to highlight the impact of the "wrong" procedure based on a single step approach (i.e. not treating donor as ungauged). In our opinion, it helps to understand two important choices we made: treating the donor catchments as ungauged and using different values for different flow groups. The objective of this test will be clarified.

- Done

3. I think there needs to be some additional strategies for validation of the uncertainty estimates. I would also ask the authors to consider other behaviors typical of confidence or predication estimates and test whether their approach follows what would be expected behavior, such as the effect of sample size or

changes in the estimates related to different flow categories. Is there null hypothesis for the method that could be tested?

We believe that our evaluation of uncertainty estimates based on three expected qualities follows common practice and is deemed sufficient to support the key points of the paper. We do not believe that testing uncertainty estimates from different perspective can be framed into a null hypothesis. However, testing reliability is essential and the coverage ratio we used provides a way to investigate if the method is able to yield reliable estimates.

- No modification introduced

4. Please provide more details in the text for Section 5.3. The use of groups seems to be somewhat arbitrary and the authors should expand more on their findings here. What would the authors recommend for a practitioner trying to use this approach?

We agree that the choice of 10 groups appears quite arbitrary. Our main motivation was to account for potential changes across different flow groups, but this has to be balanced with making sure that the number of points inside each group is sufficient to obtain reliable estimates of the empirical quantiles. We will expand more on this issue in the revised paper.

- Done

Minor comments:

p. 8045, line 22: Change to read "Here we consider a target ungauged catchments (TUC)…"

Thanks.

- Done

p. 8045, lines 23-26: The subscripts and superscripts seem inconsistent to me. For any one ungauged catchment, the authors define its neighbors as NGC_1, NGC_2, *etc. I think that would mean that in the next sentence, the subscripts should stay the same and the superscript should be i's. Maybe it woud be better to say something like, "For the ith ungauged catchments, there are n neighbouring catchments with the notation:* NGC_{1i}, NGC_{2i}, NGC_{3i}, *etc.*

Thanks. We will make appropriate changes to make it easier to understand.

- Done

p. 8046, line 13: Think it should be "error" and not "errors"

Thanks.

- Done

ANONYMOUS REFEREE #5 (R5)

The paper presents an interesting approach allowing for assessing uncertainty of flow estimates in ungauged catchments. It is well motivated, refers to the relevant sources and well structured. Illustrative material is adequate. It is a very welcome addition to the PUB, and at the same time to the uncertainty-related studies. It can be recommended to publication provided the comments below are addressed.

We thank reviewer R5 for his positive comments about the paper. We will take his comments into consideration to enhance the revised manuscript.

This review is one of the last submitted, so I can be brief since a number of points raised by other reviewers I share as well. However there are couple of additional points that are worth stressing, and which are recommended to address in the revision.

I would define the notion of the total uncertainty clearer pointing at the main source of it. The problem is that in some earlier studies the 'total' and 'residual' uncertainty are sometimes used interchangingly so some clarity in definitions is needed ('total' may be treated as including all possible sources of uncertainty (e.g. including input) which is not the case here).

Agreed. This point was also raised by another reviewer. We will add a paragraph at the beginning of Section 3 to clarify our approach, and we will change the expression "model uncertainty" into "global uncertainty".

- Done

The paper is very concise but not always easy to understand due to lack of formal representation of ideas; I would introduce more formalism in describing the main procedure on pp 8045-8046, e.g. use some notations for flows for catchments NGC, groups, multiplicative coefficients, etc. This is easy to do.

Agreed. We will introduce more formalism, write new equations, rewrite the equations that were not clear and also add new figures to help understanding the approach and the evaluation strategy.

- We introduced a general diagram of the method that intends to clarify the various steps and modified some notations in the description of the method. But we did not feel that it was necessary to introduce new equations to define the multiplicative coefficients or the groups.

Some more clarity and rigour may be needed in the statements like:

8046, L7: The groups are based on the quantiles of the simulated discharges, so that each group is equally populated. L8: The subdivision into flow groups allows accounting for the heteroscedasticity of model errors. L11: Put together the relative errors from the donors according to the group they belong to.

Agreed.

- Modified

On p 8050 (Sec. 5) the reader may find more explanation of the methodology but it comes a bit late; I would be clearer in the description of the methodology in Section 3, I think this is an important point to address.

Agreed. We will make it clear in Section 3.

- Done

P 8046: groups: would they be better described as intervals?

We do not believe that the groups will be better described as intervals because the groups are defined based on the quantiles of the empirical distribution of the simulated discharge values and not based on absolute values.

- No modification made

The presented methodology contains couple of elements that may require somewhat stronger justification, e.g. creating 10 groups, using multiplicative coeffs.

Agreed. The choice of the number of groups has to be better explained, and the use of multiplicative coefficients has to be justified. We used multiplicative coefficients instead of additive coefficients because it is the easiest way to make sure that the prediction bounds are positive. And we used 10 groups because we had to balance two objectives: having a sufficient number of points inside each groups and describing how the multiplicative coefficients vary with the magnitude of the simulated discharge. We will add a few sentences to discuss the mentioned choices.

- Done

907 catchments is great to have, but I suppose many readers would like to read about the recommendations on using this method in less data-rich cases.

Agreed. We will mention this limitation of our work.

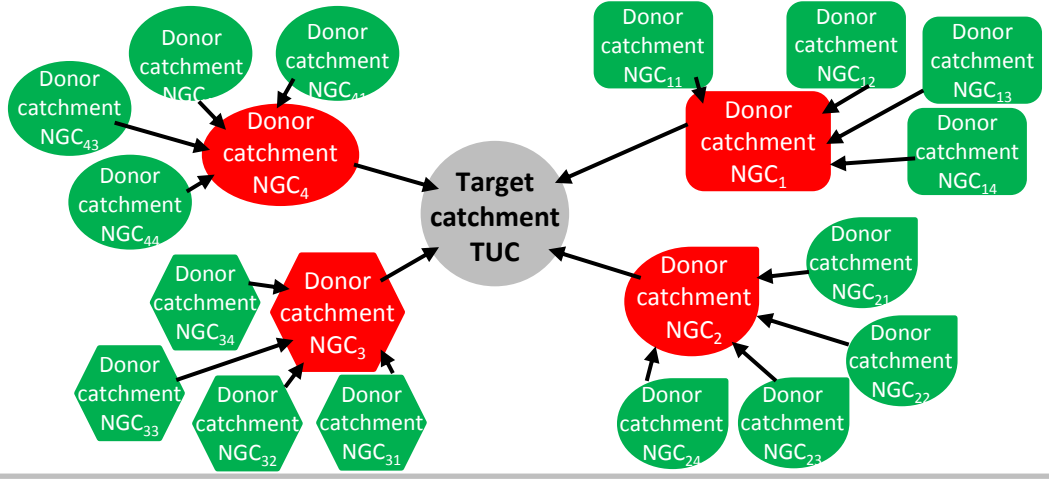
- See the complementary analysis.

In the version for printing most figures are hardly readable, it is suggested to check this.

Thank you very much for noticing this issue. We will make the appropriate modifications to have better figures.

- We include the figures in pdf format.

Step 1: Selection of catchments

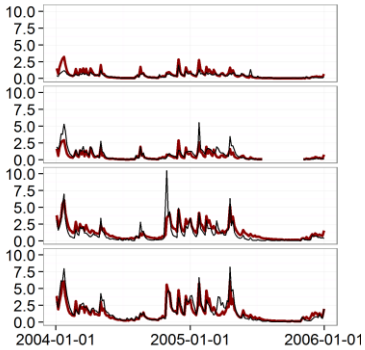


Step 2: Regionalisation applied to NGC_i

2a)

Simulations of discharges for NGC_i catchments treated as ungauged

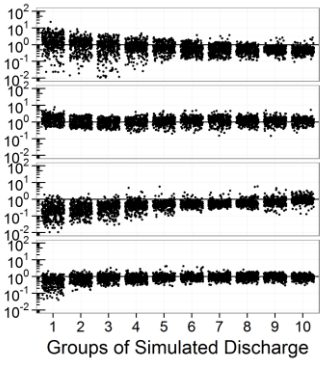
Parameter sets from NGC_{4j} NGC_{3j} NGC_{2j} NGC_{1j} with $j \in [1;4]$



NGC_1 NGC_2 NGC_3 NGC_4

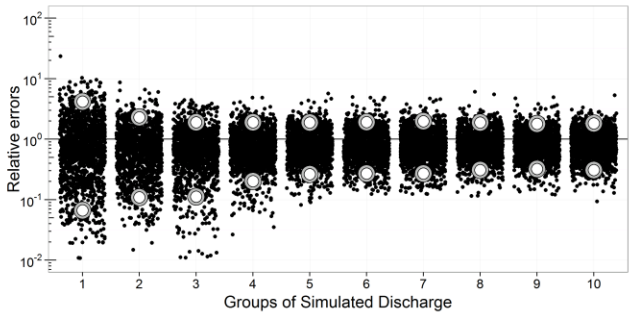
2b)

Plots of relative discharge errors by flow groups for NGC_i



Step 3: Computation of multiplicative coefficients

Agregation of the relative errors for NGC_i and computation of empirical quantiles for each flow group

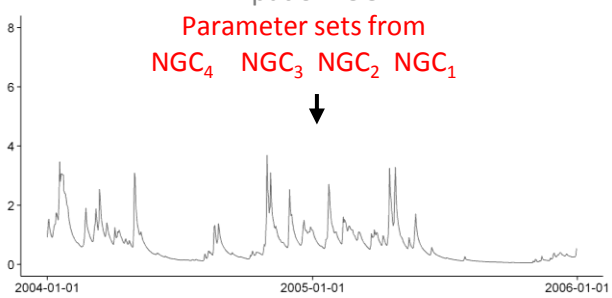


Step 4: Computation of uncertainty bounds for TUC

4a)

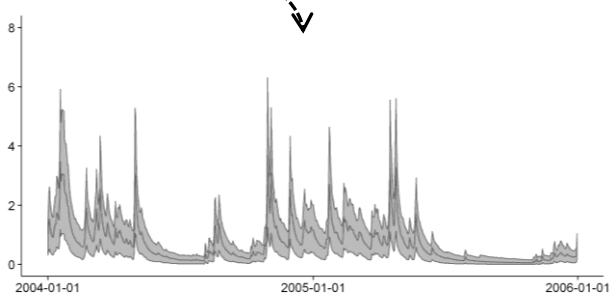
Input of TUC

Parameter sets from NGC_4 NGC_3 NGC_2 NGC_1

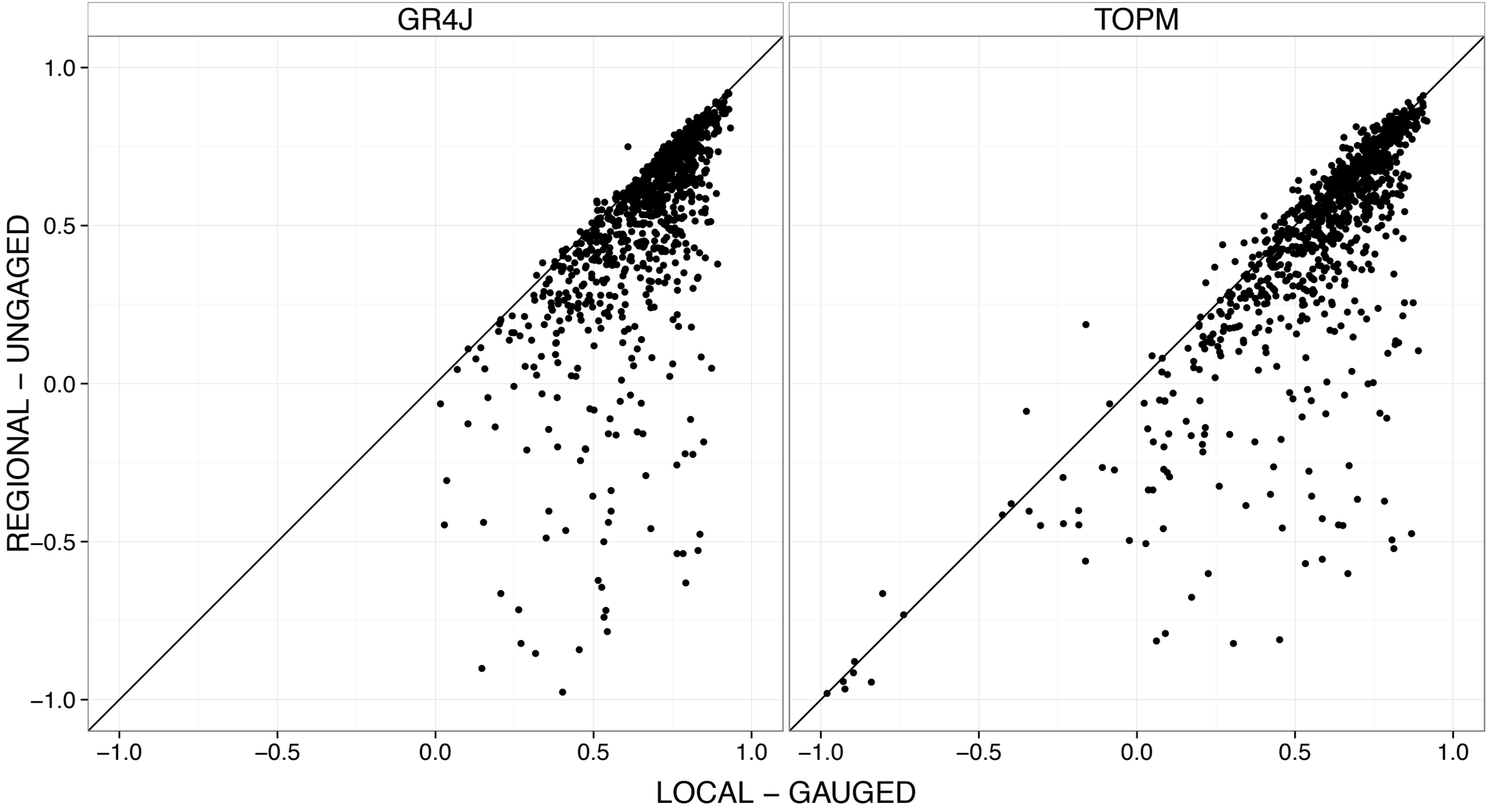


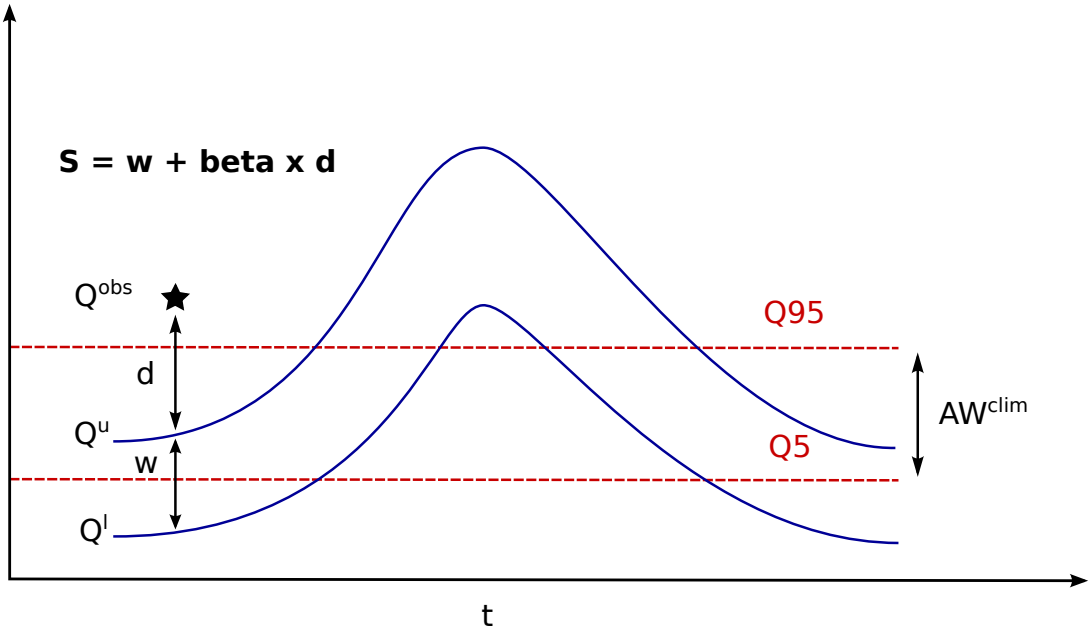
Simulations of TUC discharges

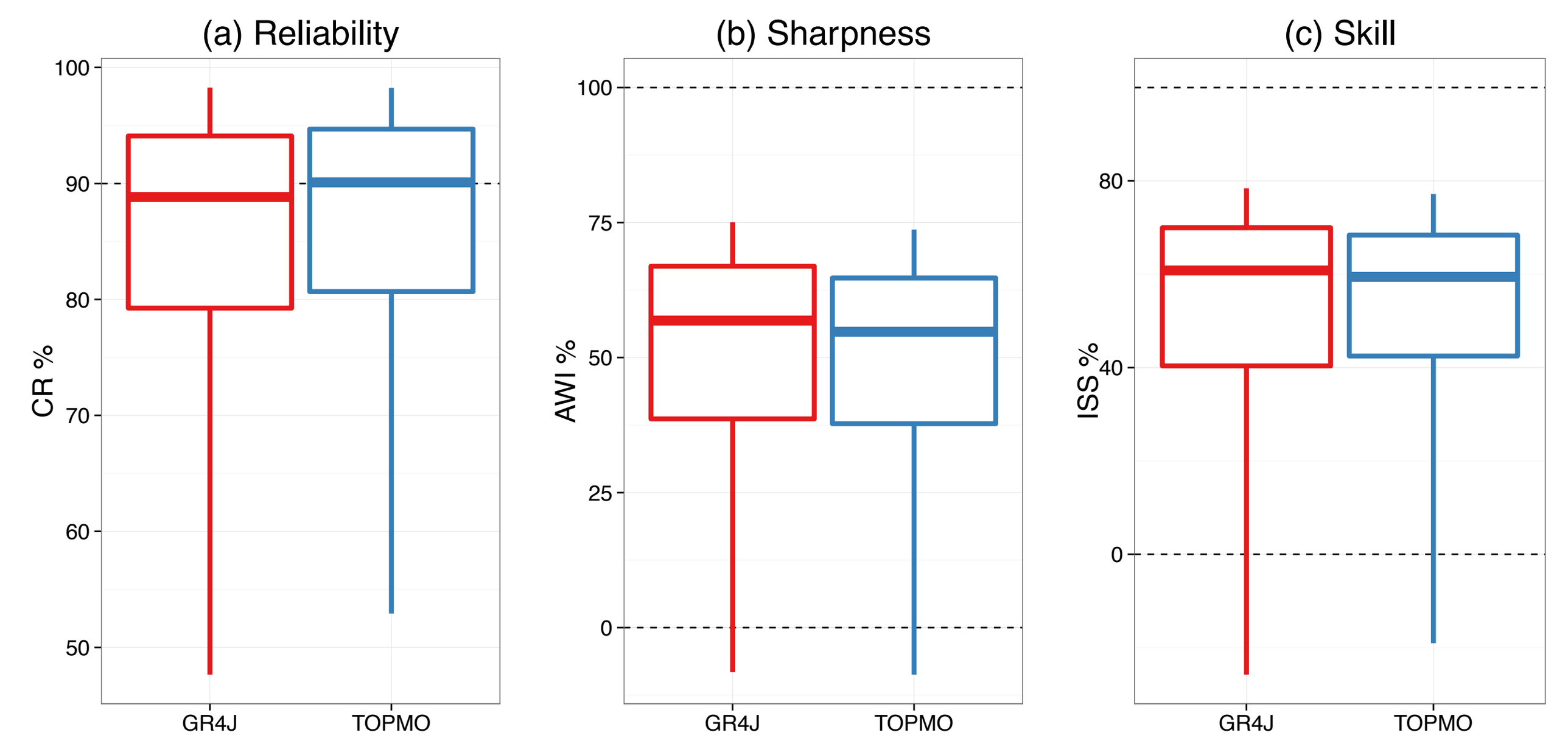
4b)

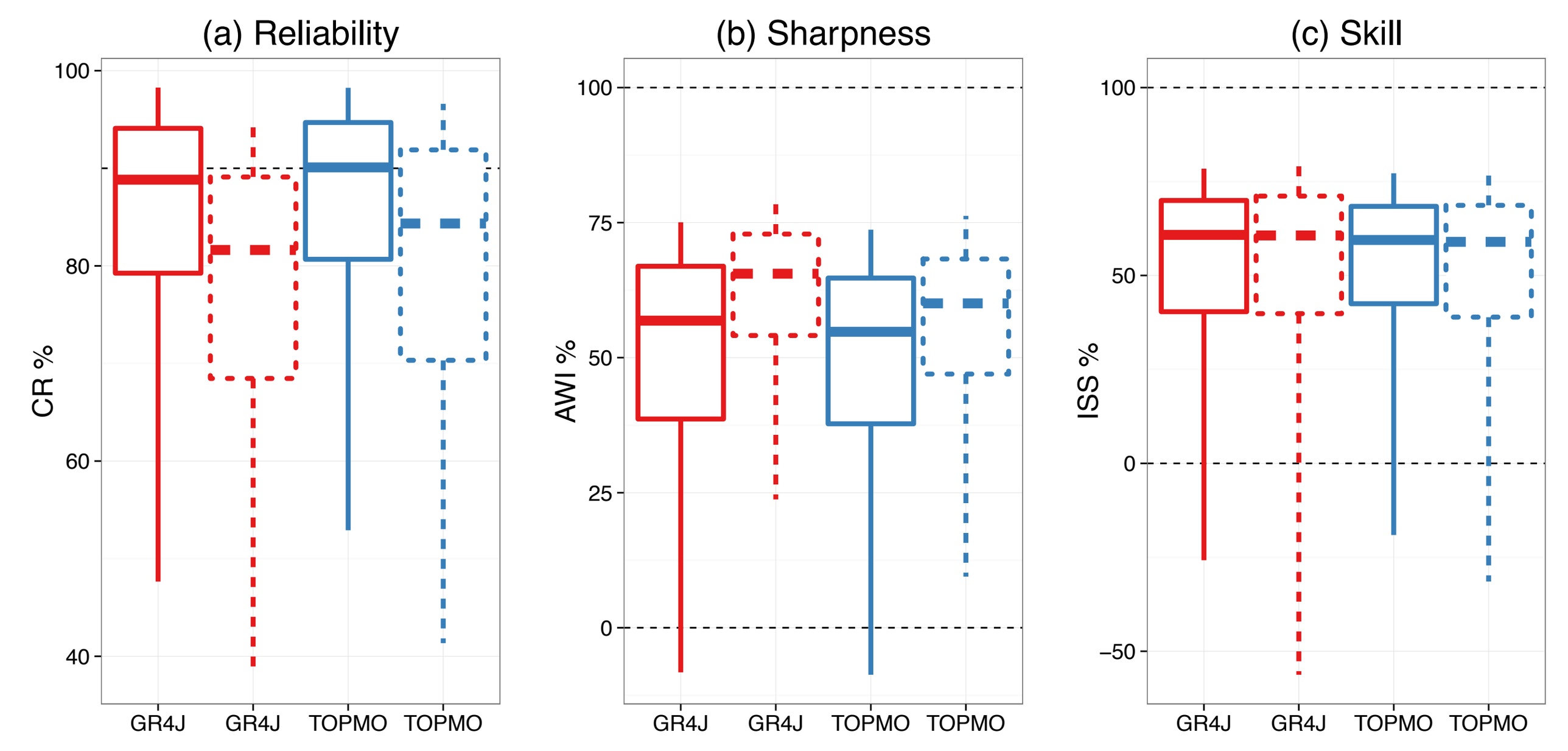


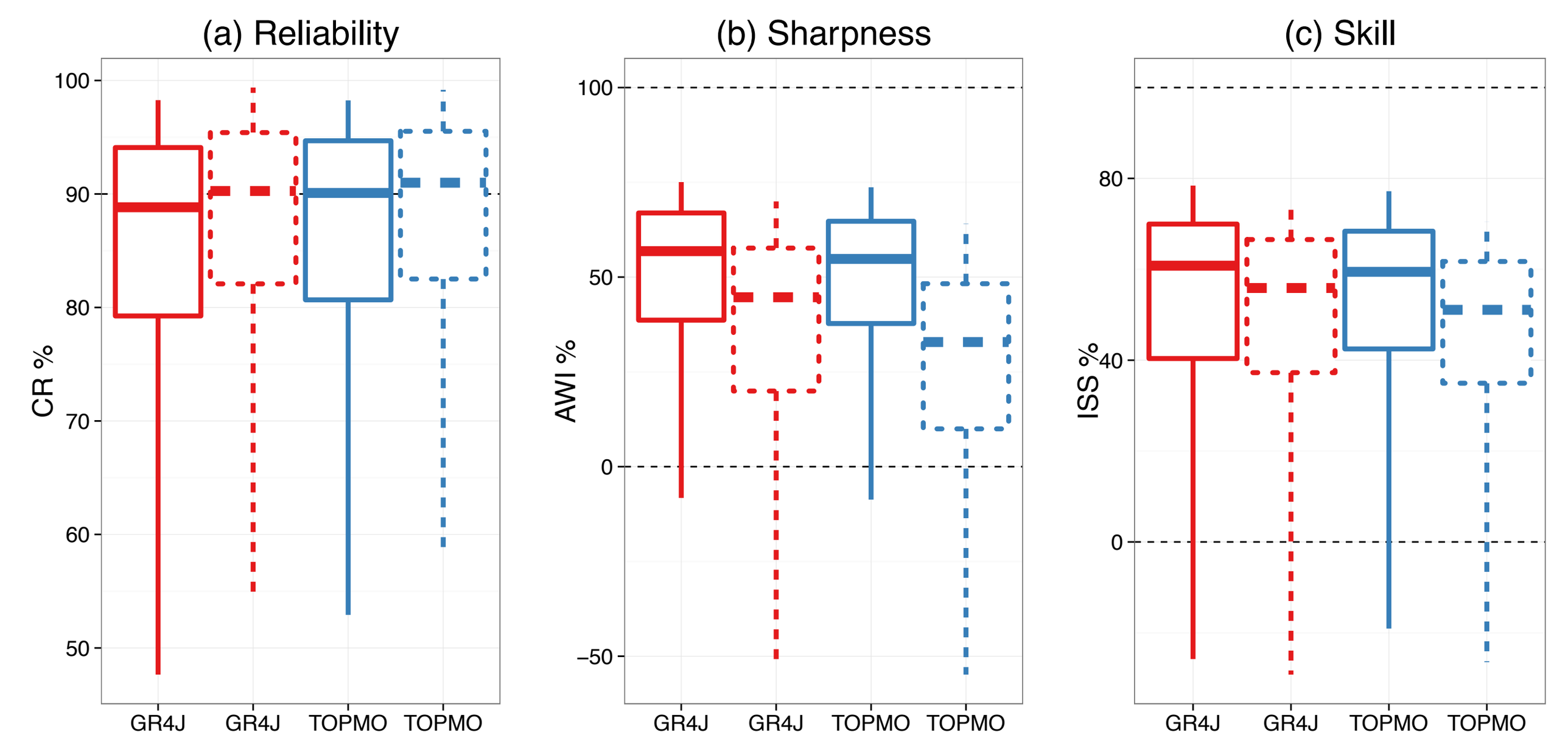
Simulated TUC discharges with associated uncertainty bounds

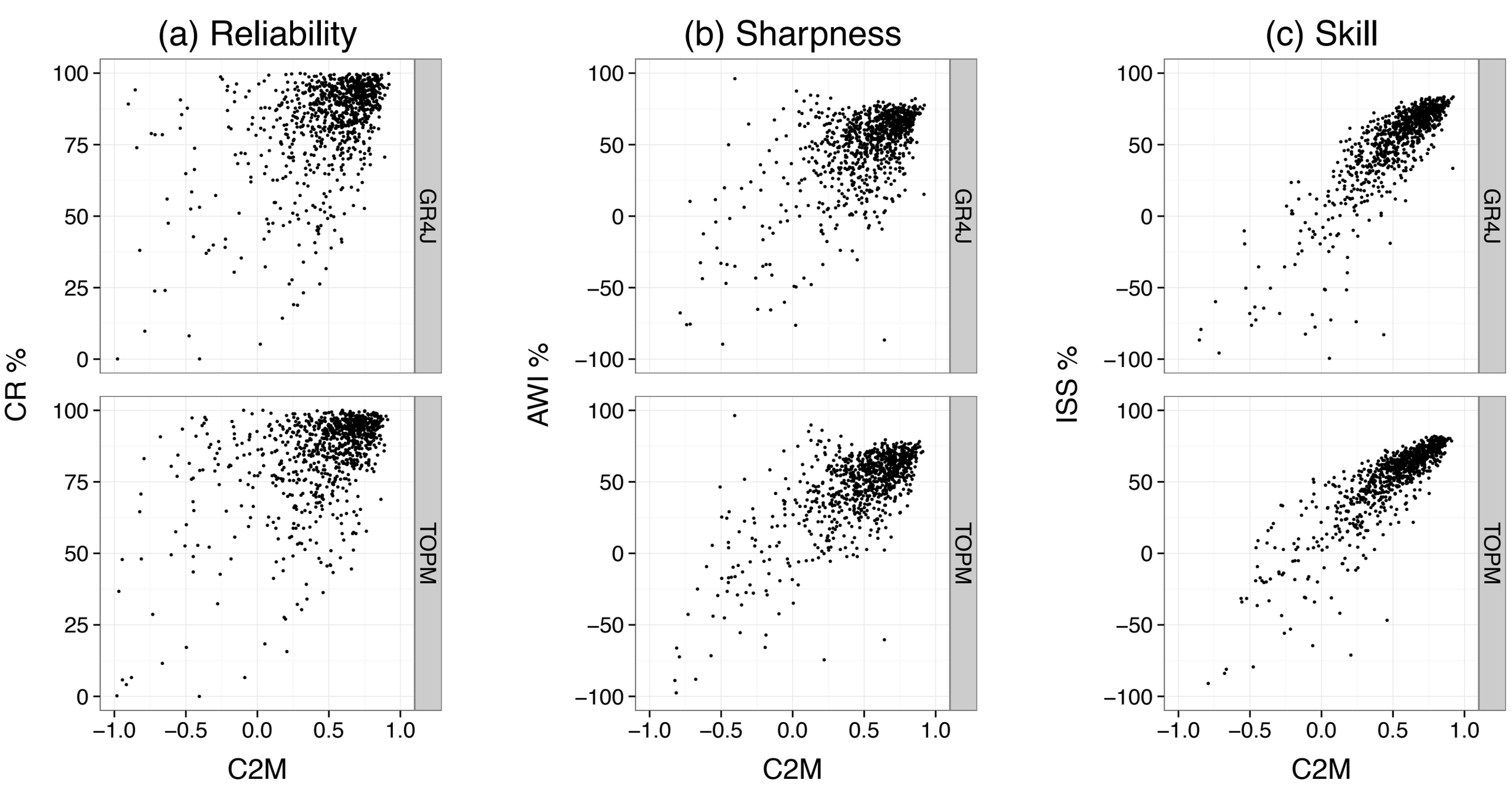

GR4J
TOPM
REGIONAL – UNGAGED
LOCAL – GAUGED
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

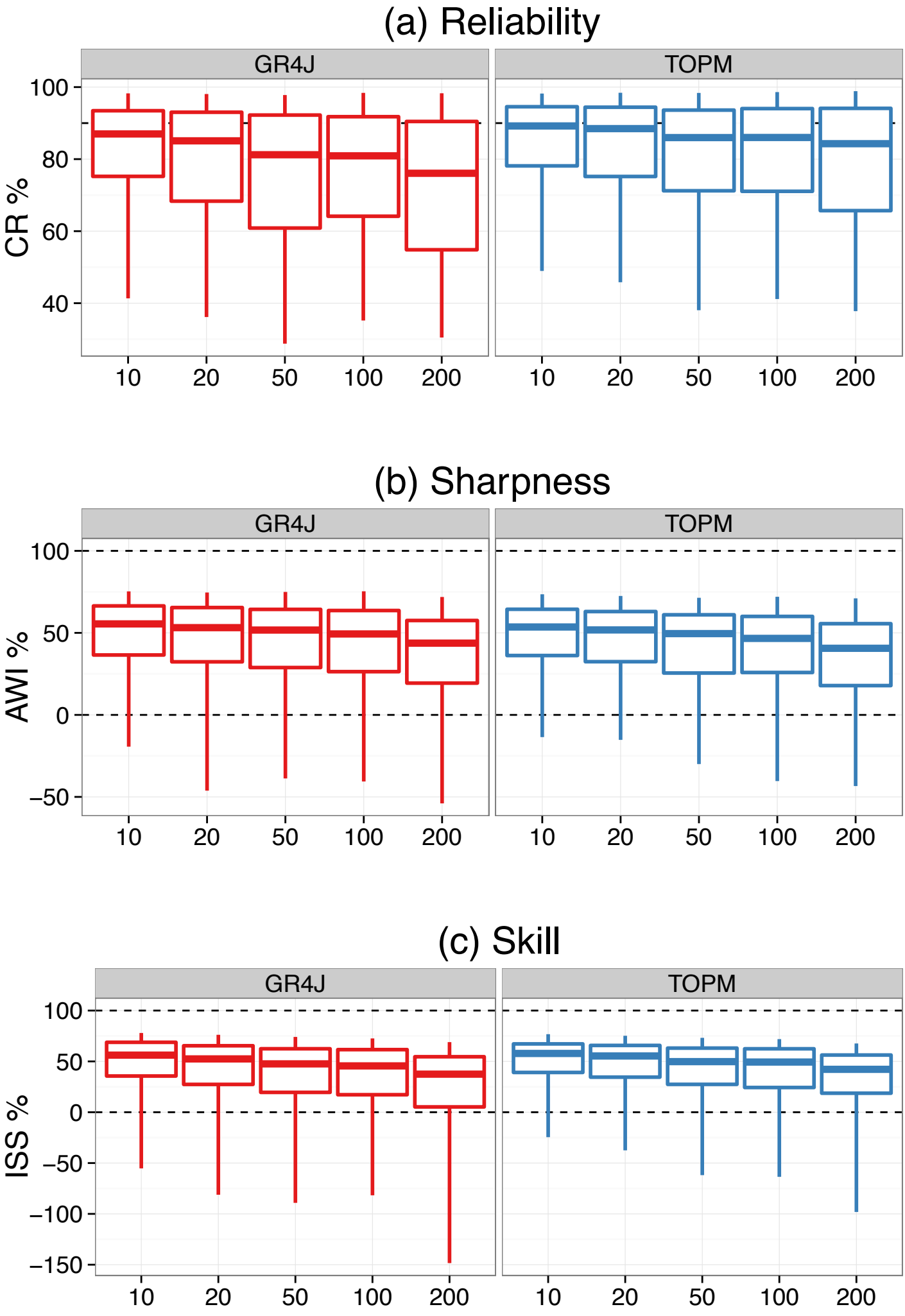

S = w + beta x d
Q^obs
d
Q^u
w
Q^l
Q95
Q5
AW^clim
t


(a) Reliability
CR %
100
90
80
70
60
50
GR4J
TOPMO
(b) Sharpness
AWI %
100
75
50
25
0
GR4J
TOPMO
(c) Skill
ISS %
80
40
0
GR4J
TOPMO


(a) Reliability
CR %
100
80
60
40
GR4J
GR4J
TOPMO
TOPMO
(b) Sharpness
AWI %
100
75
50
25
0
GR4J
GR4J
TOPMO
TOPMO
(c) Skill
ISS %
100
50
0
−50
GR4J
GR4J
TOPMO
TOPMO


(a) Reliability
CR %
100
90
80
70
60
50
GR4J
GR4J
TOPMO
TOPMO
(b) Sharpness
AWI %
100
50
0
−50
GR4J
GR4J
TOPMO
TOPMO
(c) Skill
ISS %
80
40
0
GR4J
GR4J
TOPMO
TOPMO


(a) Reliability
(b) Sharpness
(c) Skill
CR %
AWI %
ISS %
C2M
GR4J
TOPM
100
75
50
25
0
−50
−100
−1.0
−0.5
0.0
0.5
1.0


(a) Reliability
GR4J
TOPM
CR %
100
80
60
40
10
20
50
100
200
(b) Sharpness
GR4J
TOPM
AWI %
100
50
0
−50
10
20
50
100
200
(c) Skill
GR4J
TOPM
ISS %
100
50
0
−50
−100
−150
10
20
50
100
200

Manuscript prepared for Hydrol. Earth Syst. Sci. Discuss.
with version 2014/09/16 7.15 Copernicus papers of the LaTeX class copernicus.cls.
Date: 31 January 2015

Transferring ~~model~~ global uncertainty estimates from gauged to ungauged catchments

F. Bourgin[1], V. Andréassian[1], C. Perrin[1], and L. Oudin[2]

[1]Irstea, UR HBAN, 1 rue Pierre-Gilles de Gennes, CS 10030, 92761 Antony Cedex, France
[2]UPMC Univ Paris 06, UMR 7619 Metis, Case 105, 4 place Jussieu, 75005 Paris, France

Correspondence to: F. Bourgin (francois.bourgin@irstea.fr)

Abstract

Predicting streamflow hydrographs in ungauged catchments is a challenging issue, and accompanying the estimates with realistic uncertainty bounds is an even more complex task. In this paper, we present a method to transfer ~~model~~ global uncertainty estimates from gauged to ungauged catchments and we test it over a set of 907 catchments located in France, using two rainfall–runoff models. We evaluate the quality of the uncertainty estimates based on three expected qualities: reliability, sharpness, and overall skill. The robustness of the method to the availability of information on gauged catchments was also evaluated using a hydrometrical desert approach. Our results show that the method holds interesting perspectives, providing in ~~most~~ a majority of cases reliable and sharp uncertainty bounds at ungauged locations.

1 Introduction

1.1 Predicting streamflow in ungauged catchments with uncertainty estimates

Predicting the entire runoff hydrograph in ungauged catchments is a challenging issue that has attracted much attention during the last decade. In this context, the use of suitable conceptual rainfall–runoff models has proved to be useful, and because traditional calibration approaches based on observed discharge data cannot be applied in ungauged catchments, other approaches are required. Various methods have been proposed for the estimation of rainfall–runoff model parameters in ungauged catchments, as reported by the recent synthesis of the Prediction in Ungauged Basins (PUB) decade (Blöschl et al., 2013; Hrachowitz et al., 2013; Parajka et al., 2013).

The estimation of predictive uncertainty is deemed good practice in any environmental modelling activity (Refsgaard et al., 2007). In hydrological modelling, the topic has been widely discussed for years, and there is still no general agreement about how to adequately quantify uncertainty. In practice, various methodologies are currently used.

For gauged catchments, the methodologies include Bayesian ~~calibration and prediction~~ approaches (see e.g., the review of Liu and Gupta, 2007), informal methods related to the GLUE framework (Beven and Freer, 2001), multi-model approaches (Duan et al., 2007; Velazquez et al., 2010) and other ~~total~~ global uncertainty quantification methods (Montanari and Brath, 2004; Solomatine and Shrestha, 2009; Weerts et al., 2011; Ewen and O'Donnell, 2012). A comprehensive review of the topic can be found in Matott et al. (2009) and Montanari (2011).

While many methods have been proposed for gauged catchments, only a few have been proposed for the estimation of predictive uncertainty on ungauged catchments. McIntyre et al. (2005) presented a GLUE-type approach consisting of transferring ensembles of parameter sets obtained on donor (gauged) catchments to target (ungauged) catchments. More recently, a framework based on constrained parameter sets was applied in several studies ~~(Yadav et al., 2007; Zhang et al., 2008; Winsemius et al., 2009; Bulygina et al., 2011, 2012; Kapa~~ It is a two-step procedure. The first step consists in estimating with uncertainty various summary metrics of the ~~hydrographs~~hydrograph, also called "signatures" of the catchments, or gathering other "soft" or "hard" information at the target ungauged catchment. The second step is the selection of an ensemble of model parameter sets~~. "Acceptable~~: "acceptable" or "behavioural" parameter sets are those that yield sufficiently close simulated summary metrics compared to ~~regionalized metrics. The~~ the regionalized metrics obtained during the first step. A bayesian approach can also be used (Bulygina et al., 2011, 2012) . The parameter sets are given a relative weight depending on the proximity of their summary metrics compared to regionalized metrics and depending on a priori information. The reader can refer to Wagener and Montanari (2011) for a comprehensive description of both formal and informal methods belonging to this framework.

One difficulty of the above mentioned approaches lies in the interpretation of the uncertainty bounds obtained from the parameter ensemble predictions. As noted by McIntyre et al. (2005) and Winsemius et al. (2009), the uncertainty bounds cannot easily be interpreted as confidence intervals, and thus it remains difficult to use them in practice. In

addition, solely relying on an ensemble of model parameter sets to quantify total predictive uncertainty is often not sufficient when imperfect rainfall–runoff models are used.

A pragmatic alternative consists in addressing separately the parameter estimation and the global uncertainty estimation issues. It has been argued by several authors (Montanari and Brath, 2004; Andréassian et al., 2007; Ewen and O'Donnell, 2012) that a posteriori characterization of modelling errors of a "best" or "optimal" simulation can yield valid uncertainty bounds at gauged locations. In earlier studies, the terms of total uncertainty, global uncertainty or post-processing approach have been used interchangeably to refer to this approach. The various sources of uncertainty are indeed lumped into an unique error term with the goal to estimate uncertainty bounds for model outputs.

As stated by Solomatine and Shrestha (2009),

> The historical model residuals (errors) between the model prediction $\hat{y}$ and the observed data y are the best available quantitative indicators of the discrepancy between the model and the real-world system or process, and they provide valuable information that can be used to assess the predictive uncertainty.

Similarly, one could argue that the model residuals between the model prediction and the observed data at *neighbouring gauged locations* are, perhaps, the best available indicators of the discrepancy between the model and the real-world system at *the target ungauged location*, under the condition that the increase of uncertainty introduced by the regionalisation step compared to the calibration step is adequately taken into account.

The only attempt we are aware of to apply a ~~total~~ global uncertainty estimation approach at ungauged location is the one presented by Roscoe et al. (2012). They quantified uncertainty for river stage prediction at ungauged locations by first ~~interpolating~~ estimating the residual errors at ungauged locations based on residual errors at gauged locations, and then applying quantile regression to these estimated errors.

1.2 Scope of the paper

The aim of this paper is to provide an estimation of the ~~total~~ global uncertainty affecting runoff prediction at ungauged locations when a rainfall–runoff model and a regionalisation scheme ~~is~~ are used.

To our knowledge, a framework based on residual errors and ~~total~~ global uncertainty quantification has not yet been extensively tested in the context of prediction in ungauged catchments. This paper contributes to the search for methods able to provide uncertainty estimates when runoff predictions in ungauged catchments are sought.

2 Data and methods

Our objective is not to develop a new parameter regionalisation approach. Therefore, we purposely chose to use ready-to-use materials and methods and only focus on the uncertainty quantification issue. This study can be considered as a follow-up of the work ~~made~~ by Oudin et al. (2008) on the comparison of regionalisation approaches. We only provide here an overview of the data set, the rainfall–runoff models and the parameter calibration and regionalisation approach, since the details can be found in Oudin et al. (2008).

2.1 Data set

A database of 907 French catchments was used. They represent various hydrological conditions, given the variability in climate, topography, and geology in France. This set includes fast responding Mediterranean catchments with intense precipitation as well as larger, ~~groundwater-dominated~~ groundwater–dominated catchments. Some characteristics of the data set are given in Table 1. Catchments were selected to have limited snow influence, since no snowmelt module was used in the hydrological modelling. Daily rainfall, runoff, and potential evapotranspiration (PE) data series over the 1995–2005 period were available. Meteorological inputs originate from Météo-France SAFRAN reanalysis (Vidal et al., 2010). PE was estimated using the temperature-based formula proposed by

Oudin et al. (2005). Hydrological data were extracted from the HYDRO national archive (www.hydro.eaufrance.fr).

2.2 Rainfall–runoff models

Two daily, continuous lumped rainfall–runoff models were used:

- The GR4J rainfall–runoff model, an efficient and parsimonious daily lumped continuous rainfall–runoff model described by Perrin et al. (2003).
- The TOPMO rainfall–runoff model, inspired by TOPMODEL (Beven and Kirkby, 1979). This version was tested on large data sets and showed performance comparable to that of the GR4J model, while being quite different (Michel et al., 2003; Oudin et al., 2008, 2010).

Using these two models rather than a single one makes it possible to draw more general conclusions. The two models use a soil moisture accounting procedure as well as routing stores. However, their differ markedly in the formulation of their functions. While the GR4J model uses two non-linear stores and a unit-hydrograph, the TOPM model uses a linear and an exponential stores, and a pure time delay.

The GR4J and TOPMO models have four and six free parameters respectively. On gauged catchments, parameter estimation is performed using a local gradient search procedure, applied in combination with a pre-screening of the parameter space (Mathevet, 2005; Perrin et al., 2008). This optimization procedure has proved to be efficient in past applications for the conceptual models used here. As objective function, we used the Nash and Sutcliffe (1970) criterion computed on root square transformed flows (NSVQ). This criterion was shown to yield a good compromise between different objectives (Oudin et al., 2006).

2.3 Regionalisation approach

By definition, no discharge data ~~is~~ are available for calibrating parameter sets at ungauged ~~location~~locations. Thus, other strategies are needed to estimate the parameters of hydrological models for prediction in ungauged catchments.

Oudin et al. (2008) assessed the relative performance of three classical regionalisation schemes over a set of French catchments: spatial proximity, physical similarity and regression. Several options within each regionalisation approach were tested and compared. Based on their results, the following choices were made here for the regionalisation approach, as they offered the best regionalisation solution:

- Poorly modelled catchments were discarded as potential donors: only catchments with a performance criterion NSVQ in calibration above 0.7 were used as possible donors.
- The spatial proximity approach was used. It consists of transferring parameter sets from neighbouring catchments to the target ungauged catchment. Proximity of the ungauged catchments to the gauged ones was quantified by the distances between catchments centroids.
- The output averaging option was chosen. It consists of computing the mean of the streamflow simulations obtained on the ungauged catchment with the set of parameters of the donor catchments.
- The number of neighbours was set to 4 and 7 catchments for GR4J and TOPMO respectively, following the work by Oudin et al. (2008) .

3 Proposed approach: transfer of relative errors by flow groups

3.1 Description of the method

Transferring calibrated model parameters from gauged catchments to ungauged catchment is a well established approach when parameters cannot be inferred from available data.

The method presented here extends the parameter transfer approach to the domain of uncertainty estimation.

The main idea underlying the proposed approach is (i) to treat each donor as if it was ungauged (simulating flow though the above described regionalisation approach), (ii) characterize the empirical distribution of relative errors for each of these donors, and (iii) transfer ~~model~~ global uncertainty estimates to the ungauged catchment.

The methodology used to transfer ~~model~~ global uncertainty estimates can be described by the following steps, illustrated by ~~Figs~~Fig. 1~~to **??**~~:

1. Selection of catchments
   Here we consider a target ~~catchment as ungauged, called TUC~~ungaged catchment (TUC). This catchment has n neighbouring gauged catchments, called NGC_1, $\text{NGC}_2, \ldots, \text{NGC}_n$. ~~If the NGC_i catchment was now considered ungauged, one could also consider~~ For the ith catchment NGC_i, there are n neighbouring ~~catchments, called $\text{NGC}_i^1, \text{NGC}_i^2, \ldots, \text{NGC}_i^n$~~catchment with the notation: NGC_{i1}, $\text{NGC}_{i2}, \ldots, \text{NGC}_{in}$. Obviously, the TUC catchment would be excluded from this set of second order donor catchments.

2. Application of the parameter regionalisation scheme to the donor catchments NGC_i

   a. Apply the parameter regionalisation scheme to obtain a simulated discharge time series for each NGC_i using neighbours ~~NGC_i^j~~NGC_{ij} (with j between 1 and n).

   b. Compute the relative errors of ~~streamflow reconstitution~~ discharge reconstitution by comparing simulated and observed discharge series for catchment NGC_i, and create 10 groups of relative errors according to the magnitude of the simulated discharge. ~~The groups are based on the quantiles of the simulated discharges, so that each group is equally populated. The subdivision into~~ To ensure that each group contains the same number of points, the simulated discharge variable is cut into quantile groups. Using several flow groups allows ~~accounting for the heteroscedasticity~~ taking into account the possible variability of model errors characteristics.

3. Computation of the multiplicative coefficients

    a. Put together the relative errors from the donors according to the group they belong to~~.~~, i.e. for a group k, all relative errors of groups k of the n donors are assembled.

    b. Compute the empirical quantiles of the relative ~~errors~~ error distribution within each group. Each quantile of relative error can be considered a multiplicative coefficient. These multiplicative coefficients will be used to obtain the prediction bounds.

4. Computation of the uncertainty bounds for the target ungaged catchment TUC

    a. Apply the parameter regionalisation scheme to obtain a simulated discharge time series for the target ungaged catchment TUC using the parameter sets of the ~~neighbouring catchments NGC~~$_i n$ neighbouring gauged catchments NGC_1, $\text{NGC}_2, \ldots, \text{NGC}_n$.

    b. Multiply the simulated discharge by the set of multiplicative coefficients obtained at Step 3b to obtain the uncertainty bounds. The coefficients calculated for the group k are used when the simulated discharge belongs to the group k.

~~Some of the methodological choices made here will be further discussed in Sect.~~

Note that we based our approach on multiplicative errors and not on additive errors because using multiplicative coefficients yield prediction bounds for discharge that are always positive, whereas this might not always be the case with additive errors.

Finally, we mention that the choice to use 10 groups reflects a trade-off between the number of points available to obtain reasonable estimates of empirical quantiles computed for each group and an adequate treatment of the variability of the characteristics of errors with the magnitude of simulated discharge. A larger (lower) number of groups could obviously be used if more (less) data are available.

3.2 Why donors should be considered as ungaged?

The first step deserves a brief explanation. The choice to treat donors as ungaged is related to the well-known fact that the performance of rainfall–runoff models decrease when they are applied at ungaged locations with a regionalisation scheme, compared to the case where local data are available for parameter estimation. The quadratic efficiency criterion used here is the C2M (Mathevet et al., 2006) , a bounded version of the Nash and Sutcliffe (1970) efficiency (NSE) criterion. The criterion is solely based on the simulated discharges of the deterministic rainfall–runoff and is completely independent of the application of the uncertainty method. The equations are:

$$\text{C2M} = \frac{\text{NSE}}{2 - \text{NSE}} \tag{1}$$

$$\text{NSE} = 1 - \frac{\sum_{t=1}^{T} \left(Q_t^{obs} - Q_t^{sim}\right)^2}{\sum_{t=1}^{T} \left(Q_t^{obs} - \mu_o\right)^2} \tag{2}$$

where T is the total number of time-steps, Q_t^{obs} and Q_t^{sim} are the observed and simulated discharge respectively at time-step t, and μ_o is the mean of the observed discharges. The advantage of this bounded version is to avoid large negative values which are difficult to interpret.

Figure 3 illustrates the general decrease of performance for both models on our catchment set when a regionalisation scheme is used instead of a parameter estimation based on local data. As a consequence we should expect predictive uncertainty at ungauged locations to be larger than predictive uncertainty at gauged location, i.e., when the rainfall–runoff model is calibrated with observed discharge data. That is why it is necessary to consider donors as ungaged. We will come back to this important point in Section 5.

4 Quantitative evaluation of uncertainty bounds

We assessed the relevance of the 90 % uncertainty bounds by focusing on three characteristics: reliability, sharpness and overall skill. A general introduction to probabilistic evaluation can be found in Gneiting et al. (2007) and Wilks (2011) , and in Franz and Hogue (2011) for a more hydrological perspective.

Reliability refers to the statistical consistency of the uncertainty estimation with the observation, i.e., a 90 % prediction interval is expected to contain approximately 90 % of the observations if prediction errors are adequately characterized by the uncertainty estimation. To estimate the reliability, we used the coverage ratio (CR) index, computed as the percentage of observations contained in the prediction intervals.

Sharpness refers to the concentration of predictive uncertainty. ~~We used a quantitative index based on the average width~~ The average width (AW) of the uncertainty bounds ~~.~~ is widely used to quantify sharpness,

$$\mathrm{AW} = \frac{1}{T}\sum_{t=1}^{T}\left(Q_t^u - Q_t^l\right) \tag{3}$$

where Q_t^l and Q_t^u are respectively the lower and upper bounds of the prediction interval $[Q_t^l, Q_t^u]$ at time-step t.

To ease comparison between catchments, we used the width of the 90 % ~~intervals of historical flows~~ interval [Q5, Q95],

$$\mathrm{AW}^{\mathrm{clim}} = \mathrm{Q95} - \mathrm{Q5} \tag{4}$$

where Q5 and Q95 are the 5th and 95th percentiles of the flow duration curve~~, as a benchmark. The ratio (R) between these two values provides information about the reduction of uncertainty obtained by the application of the rainfall–runoff and~~ . This value characterizes the natural variability of the ~~methodology presented here, compared to the climatology. The value $1 - R$ indicates the percentage of reduction of the average width.~~

~~We call this criterion the average width~~ flows for a given catchment and has the same unit as the average width of the uncertainty bounds. It can be viewed as the average width of the uncertainty bounds of a climatological prediction, where the uncertainty bounds are constant in time and defined by the interval [Q5, Q95]. A graphical illustration is given in Fig.2.

Comparing the two values AW and AW^{clim} leads to the following dimensionless criterion called the average with index (AWI) ~~.~~

$$AWI = 1 - \frac{AW}{AW^{clim}} \tag{5}$$

It is positive if ~~the average width is reduced~~ uncertainty obtained by the application of the rainfall–runoff model and the methodology presented here is reduced compared to the climatology, and negative otherwise.

Uncertainty bounds that are as sharp as possible and reasonably reliable are sought: indeed sharp intervals that would consistently miss the target would be misleading, while overly large intervals that would successfully cover the observations at the expense of sharpness would be of limited value for decision making.

To complete the assessment of the prediction bounds, we used the interval score (Gneiting and Raftery, 2007). The interval score (IS) accounts for both ~~reliability and sharpness and provides an overall assessment of the quality of the prediction bounds~~the width of an uncertainty bound and the position of the observed value compared to the uncertainty bound. The scoring rule of the interval score at time-step t is defined as

$$S_t = \begin{cases} \left(Q_t^u - Q_t^l\right) & \text{if } Q_t^l \leq Q_t^{obs} \leq Q_t^u \\ \left(Q_t^u - Q_t^l\right) + \frac{2}{1-\beta}\left(Q_t^l - Q_t^{obs}\right) & \text{if } Q_t^{obs} < Q_t^l \\ \left(Q_t^u - Q_t^l\right) + \frac{2}{1-\beta}\left(Q_t^{obs} - Q_t^u\right) & \text{if } Q_t^{obs} > Q_t^u \end{cases} \tag{6}$$

where Q_t^{obs} is the observed value at time-step t and β is equal to 90 % since a 90 % interval is sought here. See Fig.2 for an illustration of how S is computed.

IS is the average value of ~~S~~ S_t over the time series ~~.~~

$$\mathrm{IS} = \frac{1}{T}\sum_{t=1}^{T} S_t \tag{7}$$

To ease comparison between catchments and evaluate the skill of the prediction bounds, we used the ~~unconditional climatology~~ 90 % interval [Q5, Q95] as a benchmark, similarly to what we did for the sharpness index. The climatological prediction gives uncertainty bounds that are constant in time and defined by the interval [Q5, Q95], where Q5 and Q95 are the 5th and 95th percentiles of the flow duration curve. Thus we computed the interval skill score

$$\mathrm{ISS} = 1 - \frac{\mathrm{IS}}{\mathrm{IS}^{\mathrm{clim}}} \tag{8}$$

where $\mathrm{IS}^{\mathrm{clim}}$ is the interval score obtained with the 90 % ~~climatological~~ interval [Q5, Q95]~~(Q5 and Q95 are the 5th and 95th percentiles of the flow duration curve)~~. Using skill scores is a very common approach in probabilistic forecasting. It allows to obtain dimensionless scores, similarly to the computation of the well-known Nash and Sutcliffe (1970) efficiency (NSE) criterion for assessing deterministic performance.

The ~~skill score~~ interval skill score ISS is positive when the prediction bounds are more ~~skillfull than the climatological interval~~skilful than climatology, and negative otherwise. The best value that can be achieved is egal to 1.

5 Results and discussion

5.1 Reliability, sharpness and overall skill

Figure 4 shows the distributions of the three criteria used to evaluate the uncertainty bounds on the 907 catchments. Boxplots (5th, 25th, 50th, 75th and 95th percentiles) are used to synthesize the variety of scores over the 907 catchments of the data set.

5.1.1 Reliability

For both models, half of the catchments (from the lower quartile to the upper quartile) have CR values between 80 and 95 %. The median values are equal to 89 and 90 % for GR4J and TOPMO respectively. Since a value of 90 % is expected for 90 % prediction bounds, these results suggest that the prediction bounds are ~~, in most~~ in a majority of cases, able to reflect the magnitude of errors when predicting runoff hydrographs in ungauged catchments~~.~~, even though it is clear that the perfect value of 90 % is not reached in most cases.

The CR values fall below ~~0.7~~ 70 % for around 14 % of the catchments with GR4J, and 13 % with TOPMO, which indicates cases where the proposed approach yields predictive bounds that ~~might be~~ are clearly too narrow or biased (i.e., not well ~~centered~~ centred on the observations). Note that we did not find in the literature any guidance about how to properly evaluate the CR values. The results presented here may be used as a benchmark to comparatively assess the ranges of values that would be obtained in future studies.

5.1.2 Sharpness

Regarding sharpness, it can be seen that for GR4J, half of the catchments (from the lower quartile to the upper quartile) have AWI values between 39 and 67 %, while for TOPMO corresponding values are equal to 38 and 65 %. The median values are equal to 57 and 55 % for GR4J and TOPMO respectively. The higher the AWI values, the lower the predictive uncertainty is. Since it would be utopic to expect that no errors will be made when predicting runoff hydrographs for ungauged catchments, we ~~considered~~ could consider here uncertainty reduction values between 30 and 80 % as quite satisfactory~~.~~, even though we recognize that this statement is arbitrary since there is no widely agreed values to base our analysis on.

Note that negative values are seen for 7 % of the catchments with both GR4J and TOPMO, which indicates cases where the approach ~~yield~~ yields prediction intervals whose average width is larger than the width of the ~~historical~~ [Q5, Q95] interval (Q5 and Q95 are the 5th and 95th percentiles of the flow duration curve).

5.1.3 Overall skill

Finally, Fig. 4c shows that the predictive skill for both models is positive for most catchments~~(around 92) for both models~~. For both models, half of the catchments (from the lower quartile to the upper quartile) have ISS values between 40 and 70 %. The median values are equal to 61 and 59 % for GR4J and TOPMO respectively. While it might be argued that the unconditional climatology is not a very challenging benchmark, we consider that it is still a positive and reassuring result.

5.2 Do we need to treat the donor catchments as ungauged?

~~A~~As mentioned earlier, a critical step of the proposed approach is to apply the regionalisation scheme to obtain a simulated discharge time series for each donor catchment (Step 2a). ~~This is done because we expect that predictive uncertainty at ungauged locations is larger than predictive uncertainty at gauged location, i.e., when the rainfall–runoff is calibrated with observed discharge data.~~ To assess the impact of this methodological choice, we applied the methodology described earlier to transfer uncertainty estimates, but this time the donor catchments are treated as gauged.

Similarly to Fig. 4, Fig. 5 shows the distributions of the three criteria obtained in the two cases: whether or not the donor catchments are treated as ungauged. We can observe for both models a drop in reliability, whereas sharpness increases. This is because the relative errors are smaller when the donor catchments are treated as gauged, yielding narrower but less reliable prediction bounds for the target catchment. Interestingly, this results in skill scores that are quite similar: improvements in terms of sharpness compensate decreases in terms of reliability.

Note that reliability is generally considered as a prevailing characteristic over sharpness, since it reflects the ability of the uncertainty method to adequately reflect the magnitude of errors we might expect at locations for which prediction is done. Therefore, the benefit of treating the donor catchments as ungauged clearly appears in Fig. 5a, illustrating the theoretical argument presented in the methodological section.

5.3 Do we need to use groups of relative errors?

Another critical step of the proposed approach is to use 10 groups of relative errors. The groups are defined according to the magnitude of the simulated discharge (Step 2b). This was done to take into account the fact that the characteristics of errors usually change according to the magnitude of the simulated discharge. To assess the impact of this methodological choice, we again applied the methodology described earlier to transfer ~~model~~ global uncertainty estimates, but this time using only one group ~~is used~~ instead of 10.

Figure 6 shows the distributions of the three criteria obtained in the following two cases: whether 10 groups or only one group of relative errors are used. For both models, reliability slightly ~~increase~~increases, whereas both sharpness and skill decrease. It appears that improvements in terms of reliability are not sufficient to compensate for decreases in terms of sharpness when overall skill is considered.

While it could be argued that using only one group is the preferable option because of the slight improvement in terms of reliability, in our opinion, the improvement is not sufficiently important to compensate for the decrease in terms of uncertainty reduction and skill. We definitely prefer to maintain different flow groups.

5.4 How do the performances of the rainfall–runoff models relate to the characteristics of uncertainty bounds?

To gain insights into the possible relationships between the performance of the deterministic rainfall–runoff simulations and characteristics of the uncertainty bounds at ungauged locations, the three criteria used to characterize the uncertainty bounds are plotted in Fig. 7 as function of a quadratic efficiency ~~criteria~~ criterion for the 907 catchments~~. The quadratic efficiency criterion is~~ , the C2M defined in Eq. 1.

A trend appears between deterministic performance and characteristics of the prediction bounds at ungauged locations, for the two rainfall–runoff models. The reliability index exhibits larger variability compared to the sharpness index, and the stronger link is seen for the skill score. Reliability is relatively less affected by the poor deterministic performance of

the simulation at ungauged location because there are cases where poor performance at ~~neighbour~~ neighbouring locations leads (though the transfer of relative errors) to wide prediction bounds that are able to cover the observed values. We can also observe that skill scores and C2M scores are strongly related, which indicates that when the transfer of model parameters performs well, the transfer of ~~model~~ global uncertainty estimates performs well too.

5.5 How does the method perform in data-sparse conditions?

The results presented so far were obtained with a dense network of gauging stations. To investigate the impact of the network density on our results, we applied a demanding test called the hydrometrical desert. It consists in excluding potential donors that are closer to the target ungaged catchment than a given threshold. For example, a threshold distance of 100 km means that the closest donor catchment must be at least 100 km far from the ungaged target catchment. This test results in a notable decrease of deterministic performance, as shown in Table 2, where the mean of the C2M efficiency criterion over the 907 catchments is reported, for both models. Note that this is a more demanding test than a decrease of network density, because catchments keeps the possibility to still have close donors.

Figure 6 shows the distributions of the three criteria obtained by applying the hydrometrical desert with threshold values of 10, 20, 50, 100 and 200 km, respectively. A clear decrease appears with increasing distances. While we should expect that the sharpness of the uncertainty bounds decreases because of larger errors, and that this situation leads to a decrease of skill, the results in terms of reliability reveal the limitation of the method. With increasing distances, the method becomes less able to transfer the appropriate magnitude of the larger errors.

6 Conclusions

Runoff hydrograph prediction in ungauged catchments is notoriously difficult, and attempting to estimate the predictive uncertainty in that context is a further challenge. We proposed a method allowing the transfer of ~~model~~ global uncertainty estimates from gauged to ungauged catchments. The method extends the parameter transfer approach to the domain of global uncertainty estimation.

We evaluated the approach over a large set of 907 catchments by assessing three expected qualities of uncertainty estimates, reliability, sharpness and overall skill. We applied two different rainfall–runoff models (GR4J and TOPMO) to ensure that the presented results are not model-specific. Our results demonstrate that the method is generally able to reflect model errors at ungauged locations and provide reasonable reliability. ~~We applied two different rainfall- runoff models (GR4J and TOPMO) to ensure that the presented results are not model-specific.~~

~~Although~~ Nonetheless, the following limitations of our study can be mentioned:

1. Although the approach seems promising on average on the large catchment set we used, it is not able to adequately quantify the predictive uncertainty for some catchments and it failed in some cases.

2. The method might not perform well in in regions with sparser gauging networks than the one used here, as revealed by the application of a demanding test called the hydrometrical desert.

3. We only tested the 90 % prediction intervals, whereas the method could be applied to obtain other prediction intervals. We made this choice to keep the article as simple as possible, but further work could be done in that direction.

It is worth stressing that although we used a transfer based on spatial proximity, the approach ~~is~~ presented in this article is not only independent of the rainfall-runoff model but also of the regionalisation scheme used to obtain deterministic prediction at ungauged locations~~, and any~~. Any other similarity measure could be a basis for transferring residual

errors., including physical-based similarity measures. Accordingly, the regionalisation settings, including the output averaging option, could be adapted if deemed more appropriate.

Since we believe that uncertainty quantification has to be considered in any modelling study, further work should be devoted to the search for similarity measures that do not only perform well in allowing the transfer of parameter sets from donor to target catchments, but also allow transferring modelling error characteristics.

Last, we would like to stress that the results presented in this study are expressed in terms of dimensionless measures. As such, they can provide a basis for future comparisons when prediction in ungauged catchments with uncertainty estimates is performed.

Acknowledgements. The authors thank Météo-France for providing the meteorological data and Banque HYDRO for the hydrological data. The financial support of SCHAPI to the first author is also gratefully acknowledged.

References

Andréassian, V., Lerat, J., Loumagne, C., Mathevet, T., Michel, C., Oudin, L., and Perrin, C.: What is really undermining hydrologic science today?, Hydrological Processes, 21, 2819–2822, doi:10.1002/hyp.6854, 2007.

Beven, K. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, Journal of Hydrology, 249, 11–29, 2001.

Beven, K. J. and Kirkby, M. J.: A physically based, variable contributing area model of basin hydrology, Hydrological Sciences Bulletin, 24, 43–69, doi:10.1080/02626667909491834, 1979.

Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., and Savenije, H.: Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales, Cambridge University Press, 2013.

Bulygina, N., McIntyre, N., and Wheater, H.: Bayesian conditioning of a rainfall-runoff model for predicting flows in ungauged catchments and under land use changes, Water Resources Research, 47, W02 503, doi:10.1029/2010wr009240, 2011.

Bulygina, N., Ballard, C., McIntyre, N., O'Donnell, G., and Wheater, H.: Integrating different types of information into hydrological model parameter estimation: Application to ungauged catchments and land use scenario analysis, Water Resources Research, 48, W06 519, doi:10.1029/2011wr011207, 2012.

Duan, Q., Ajami, N. K., Gao, X., and Sorooshian, S.: Multi-model ensemble hydrologic prediction using Bayesian model averaging, Advances in Water Resources, 30, 1371–1386, doi:10.1016/j.advwatres.2006.11.014, 2007.

Ewen, J. and O'Donnell, G.: Prediction intervals for rainfall-runoff models: raw error method and split-sample validation, Hydrology Research, 43, 637–648, doi:10.2166/nh.2012.038, 2012.

Franz, K. J. and Hogue, T. S.: Evaluating uncertainty estimates in hydrologic models: borrowing measures from the forecast verification community, Hydrology and Earth System Sciences, 15, 3367–3382, doi:10.5194/hess-15-3367-2011, 2011.

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, Journal of the American Statistical Association, 102, 359–378, doi:10.1198/016214506000001437, 2007.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, Journal of the Royal Statistical Society Series B-Statistical Methodology, 69, 243–268, doi:10.1111/j.1467-9868.2007.00587.x, 2007.

Hrachowitz, M., Savenije, H. H. G., Bloeschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V., Hughes, D. A., Hut, R. W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P. A., Uhlenbrook, S., Wagener, T., Winsemius, H. C., Woods, R. A., Zehe, E., and Cudennec, C.: A decade of Predictions in Ungauged Basins (PUB)a review, Hydrological Sciences Journal-Journal Des Sciences Hydrologiques, 58, 1198–1255, doi:10.1080/02626667.2013.803183, 2013.

Kapangaziwiri, E., Hughes, D. A., and Wagener, T.: Incorporating uncertainty in hydrological predictions for gauged and ungauged basins in southern Africa, Hydrological Sciences Journal-Journal Des Sciences Hydrologiques, 57, 1000–1019, doi:10.1080/02626667.2012.690881, 2012.

Liu, Y. Q. and Gupta, H. V.: Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, Water Resources Research, 43, W07 401, doi:10.1029/2006wr005756, 2007.

Mathevet, T.: Quels modèles pluie-débit globaux au pas de temps horaire ? Développements empiriques et comparaison de modèles sur un large échantillon de bassins versants, Ph.D. thesis, Paris, 2005.

Mathevet, T., Michel, C., Andréassian, V., and Perrin, C.: A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins, IAHS-AISH Publication, pp. 211–219, 2006.

Matott, L. S., Babendreier, J. E., and Purucker, S. T.: Evaluating uncertainty in integrated environmental models: A review of concepts and tools, Water Resources Research, 45, W06 421, doi:10.1029/2008wr007301, 2009.

McIntyre, N., Lee, H., Wheater, H., Young, A., and Wagener, T.: Ensemble predictions of runoff in ungauged catchments, Water Resources Research, 41, W12 434, doi:10.1029/2005wr004289, 2005.

Michel, C., Perrin, C., and Andreassian, V.: The exponential store: a correct formulation for rainfall-runoff modelling, Hydrological Sciences Journal-Journal Des Sciences Hydrologiques, 48, 109–124, doi:10.1623/hysj.48.1.109.43484, 2003.

Montanari, A.: Uncertainty of Hydrological Predictions, in: Treatise on Water Science, edited by Peter, W., pp. 459–478, Elsevier, Oxford, 2011.

Montanari, A. and Brath, A.: A stochastic approach for assessing the uncertainty of rainfall-runoff simulations, Water Resources Research, 40, W01 106, doi:10.1029/2003wr002540, 2004.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I - A discussion of principles, Journal of Hydrology, 10, 282–290, 1970.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andreassian, V., Anctil, F., and Loumagne, C.: Which potential evapotranspiration input for a lumped rainfall-runoff model? Part 2 - Towards a simple and efficient potential evapotranspiration model for rainfall-runoff modelling, Journal of Hydrology, 303, 290–306, doi:10.1016/j.jhydrol.2004.08.026, 2005.

Oudin, L., Andreassian, V., Mathevet, T., Perrin, C., and Michel, C.: Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations, Water Resources Research, 42, W07 410, doi:10.1029/2005wr004636, 2006.

Oudin, L., Andreassian, V., Perrin, C., Michel, C., and Le Moine, N.: Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments, Water Resources Research, 44, W03 413, doi:10.1029/2007wr006240, 2008.

Oudin, L., Kay, A., Andreassian, V., and Perrin, C.: Are seemingly physically similar catchments truly hydrologically similar?, Water Resources Research, 46, W11 558, doi:10.1029/2009wr008887, 2010.

Parajka, J., Viglione, A., Rogger, M., Salinas, J. L., Sivapalan, M., and Bloeschl, G.: Comparative assessment of predictions in ungauged basins - Part 1: Runoff-hydrograph studies, Hydrology and Earth System Sciences, 17, 1783–1795, doi:10.5194/hess-17-1783-2013, 2013.

Perrin, C., Michel, C., and Andreassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275–289, doi:10.1016/s0022-1694(03)00225-7, 2003.

Perrin, C., Andreassian, V., Serna, C. R., Mathevet, T., and Le Moine, N.: Discrete parameterization of hydrological models: Evaluating the use of parameter sets libraries over 900 catchments, Water Resources Research, 44, W08 447, doi:10.1029/2007wr006579, 2008.

Refsgaard, J. C., van der Sluijs, J. P., Hojberg, A. L., and Vanrolleghem, P. A.: Uncertainty in the environmental modelling process - A framework and guidance, Environmental Modelling & Software, 22, 1543–1556, doi:10.1016/j.envost.2007.02.004, 2007.

Roscoe, K. L., Weerts, A. H., and Schroevers, M.: Estimation of the uncertainty in water level forecasts at ungauged river locations using quantile regression, International Journal of River Basin Management, 10, 383–394, doi:10.1080/15715124.2012.740483, 2012.

Solomatine, D. P. and Shrestha, D. L.: A novel method to estimate model uncertainty using machine learning techniques, Water Resources Research, 45, W00B11, doi:10.1029/2008wr006839, 2009.

Velazquez, J. A., Anctil, F., and Perrin, C.: Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments, Hydrology and Earth System Sciences, 14, 2303–2317, doi:10.5194/hess-14-2303-2010, 2010.

Vidal, J.-P., Martin, E., Franchisteguy, L., Baillon, M., and Soubeyroux, J.-M.: A 50-year high-resolution atmospheric reanalysis over France with the Safran system, International Journal of Climatology, 30, 1627–1644, doi:10.1002/joc.2003, 2010.

Wagener, T. and Montanari, A.: Convergence of approaches toward reducing uncertainty in predictions in ungauged basins, Water Resources Research, 47, W06 301, doi:10.1029/2010wr009469, 2011.

Weerts, A. H., Winsemius, H. C., and Verkade, J. S.: Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales), Hydrology and Earth System Sciences, 15, 255–265, doi:10.5194/hess-15-255-2011, 2011.

Wilks, D. S.: Statistical methods in the atmospheric sciences, Academic, Oxford, 3rd edn., 2011.

Winsemius, H. C., Schaefli, B., Montanari, A., and Savenije, H. H. G.: On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information, Water Resources Research, 45, W12 422, doi:10.1029/2009wr007706, 2009.

Yadav, M., Wagener, T., and Gupta, H.: Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, Advances in Water Resources, 30, 1756–1774, doi:10.1016/j.advwatres.2007.01.005, 2007.

Zhang, Z., Wagener, T., Reed, P., and Bhushan, R.: Reducing uncertainty in predictions in ungauged basins by combining hydrologic indices regionalization and multiobjective optimization, Water Resources Research, 44, W00B04, doi:10.1029/2008wr006833, 2008.

Table 1. Characteristics of the 907 catchments. P – precipitation, PE – potential evapotranspiration, Q – discharge.

	Percentiles				
	0.05	0.25	0.50	0.75	0.95
Catchment area (km^2)	27	73	149	356	1788
Mean annual precipitation ($mm\,yr^{-1}$)	753	853	978	1176	1665
Mean annual potential evapotranspiration ($mm\,yr^{-1}$)	549	631	659	700	772
Mean annual runoff ($mm\,yr^{-1}$)	133	233	344	526	1041
Q/P ratio	0.17	0.27	0.34	0.45	0.68
P/PE ratio	1.06	1.25	1.47	1.83	2.9
Median elevation (m)	76	149	314	645	1183

Table 2. Mean C2M over the 907 catchments of the data set, with calibration (CAL), regionalisation (REGIO), and with the hydrometrical desert (HD) defined by increasing distance 10, 20, 50, 100 and 200 km.

	CAL	REGIO	HD-10	HD-20	HD-50	HD-100	HD-200
GR4J	0,67	0,51	0,49	0,46	0,43	0,41	0,35
TOPM	0,59	0,47	0,46	0,44	0,41	0,39	0,34

~~Illustration of the proposed approach – Step 1: in **(A)**, a target catchment (grey) is considered as ungauged; this catchment has n neighbouring gauged catchments (red). In **(B)**, if one of the neighbouring catchment is now considered ungauged (green), we also consider n neighbouring catchments (yellow). Note that the target catchment is excluded from this set of second order donor catchments.~~

~~Illustration of the proposed approach – Step 2a: simulated (green, dashed) and observed (black) discharge time series for four donor catchments treated as ungauged, i.e., in which model parameters must be estimated from a regionalisation approach.~~

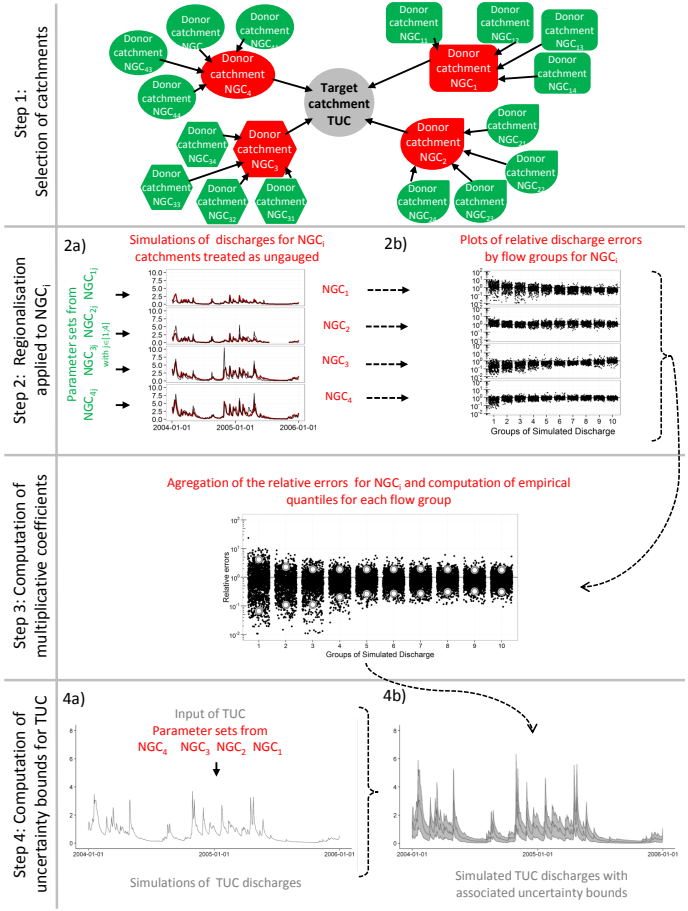

Step 1: Selection of catchments
Target catchment TUC
Donor catchment NGC1
Donor catchment NGC2
Donor catchment NGC3
Donor catchment NGC4
Donor catchment NGC5
Donor catchment NGC6
Donor catchment NGC7
Donor catchment NGC8
Donor catchment NGC9
Donor catchment NGC10
Donor catchment NGC11
Donor catchment NGC12
Donor catchment NGC13
Donor catchment NGC14
Step 2: Regionalisation applied to NGCj
2a)
Simulations of discharges for NGCi catchments treated as ungauged
Parameter sets from NGC4j NGC3j NGC2j NGC1j with j∈[1-4]
2004-01-01
2005-01-01
2006-01-01
2b)
Plots of relative discharge errors by flow groups for NGCi
Groups of Simulated Discharge
Step 3: Computation of multiplicative coefficients
Agregation of the relative errors for NGCi and computation of empirical quantiles for each flow group
Relative errors
Groups of Simulated Discharge
Step 4: Computation of uncertainty bounds for TUC
4a)
Input of TUC
Parameter sets from NGC4 NGC3 NGC2 NGC1
Simulations of TUC discharges
4b)
Simulated TUC discharges with associated uncertainty bounds

Figure 1. Illustration of the proposed approach~~– Step 2b: relative errors by flow groups; groups~~ , in the case of ~~relative errors~~ $n = 4$ donors. Red catchments are ~~defined according to~~ first-level donors while green catchments are second-level donors. See the ~~magnitude~~ text for the description of the ~~simulated discharge~~four steps.

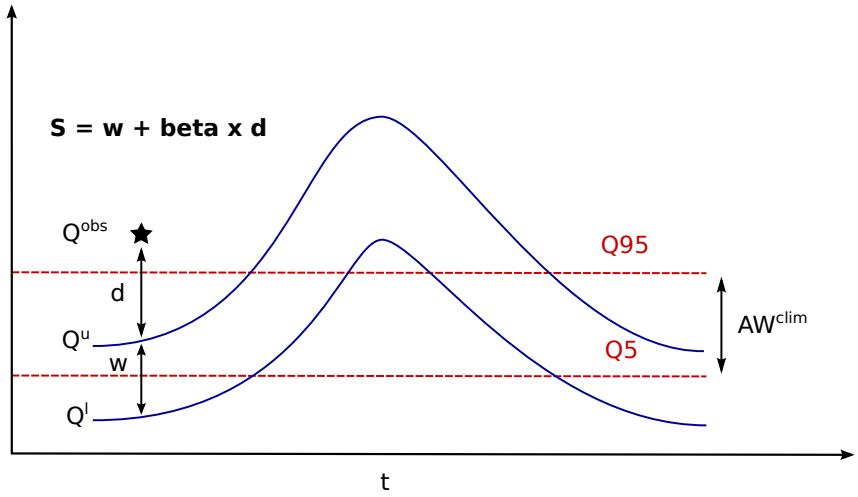


Figure 2. Illustration of the ~~proposed approach – Steps~~evaluation of the uncertainty bounds. Q5 and ~~3a~~ Q95 are the 5th and ~~b: aggregating~~ 95th percentiles of the ~~relative errors observed~~ flow duration curve. S is the interval score defined at one time-step for the ~~donors catchments; white dots correspond to~~ situation where the ~~empirical quantiles (5 and 95)~~ observed value is above the upper limit of the ~~relative errors distribution within each group~~uncertainty bound. See the text for further details.

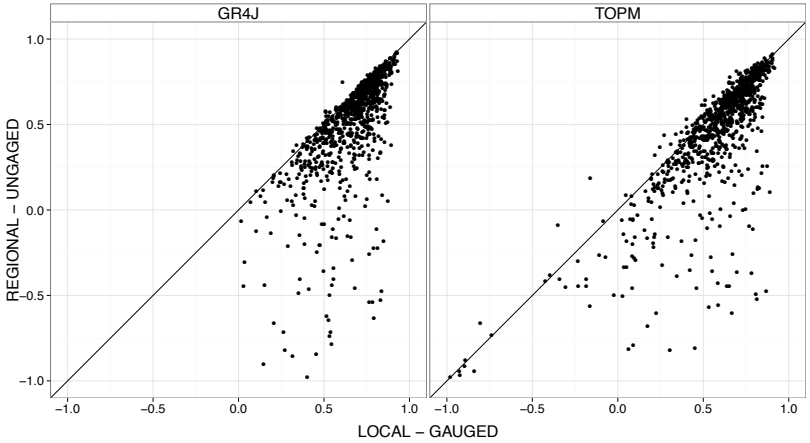


Figure 3. ~~Illustration~~ Impact of the ~~proposed approach – Steps 4a and b: simulated (red~~regionalisation scheme on deterministic performance, ~~dashed) and observed (black) discharge time series for~~ as quantified by the ~~ungauged catchments; 90uncertainty bounds~~ bounded C2M efficiency criterion. Note that in ~~grey~~a very few cases, the performance obtained with the regionalisation scheme is better than the performance obtained with calibration. This is possible because of the output averaging option used by the regionalisation scheme.

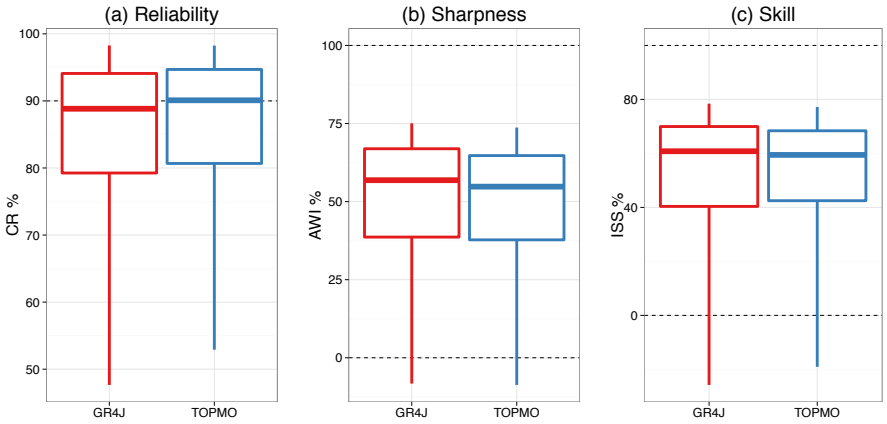


Figure 4. Distributions of the three performance criteria. Boxplots (5th, 25th, 50th, 75th and 95th percentiles) synthesize the variety of scores over the 907 catchments of the data set.

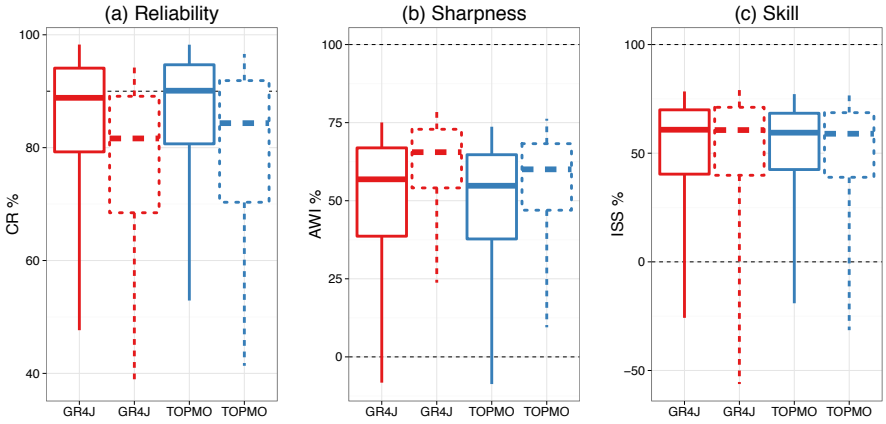


Figure 5. Distributions of the three performance criteria, obtained in two cases, (i) when the donor catchments are treated as ungauged (~~continous~~ continuous lines) and (ii) when the donor catchments are treated as gauged (dashed lines). Boxplots (5th, 25th, 50th, 75th and 95th percentiles) synthesize the variety of scores over the 907 catchments of the data set.

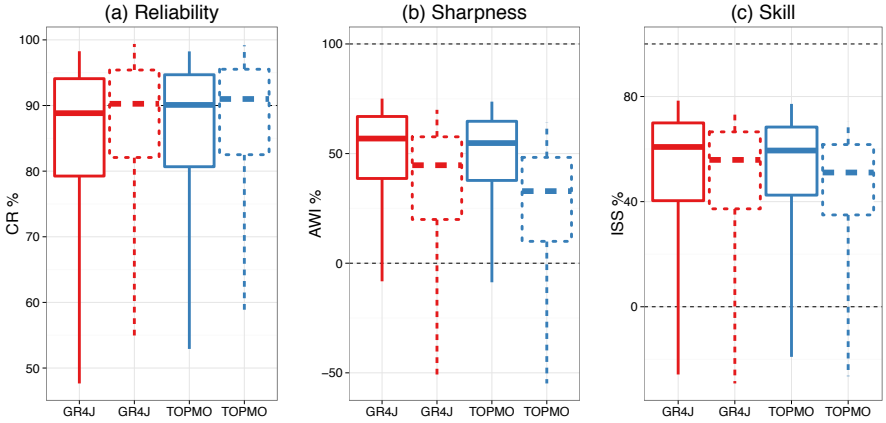


Figure 6. Distributions of the three performance criteria, obtained in two cases, (i) when 10 groups of relatives errors are used (~~continous~~ continuous lines) and (ii) when only one group is used (dashed lines). Boxplots (5th, 25th, 50th, 75th and 95th percentiles) synthesize the variety of scores over the 907 catchments of the data set.

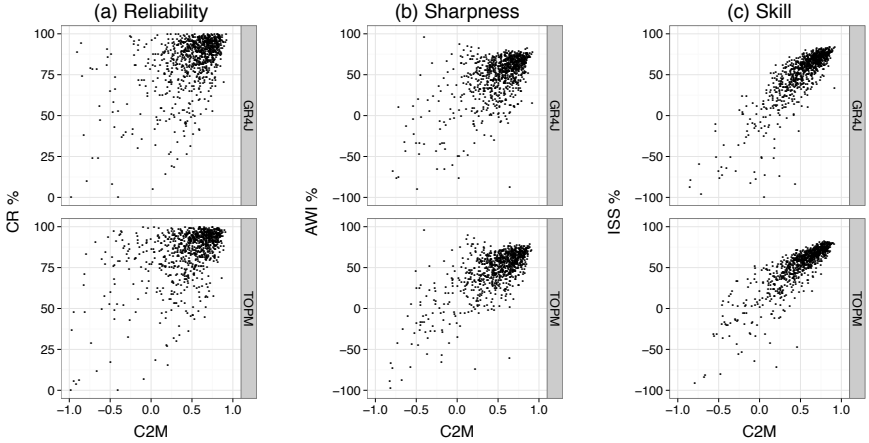


Figure 7. Impact of deterministic performance, as quantified by the bounded C2M quadratic criterion, on the three performance criteria for the 907 catchments. Note that for easing visualisation, the lower limits of AWI **(b)** and ISS **(c)** values are set to $-100\,\%$ but lower values of AWI are obtained in 7 cases for both models, and lower values are obtained in 18 and 22 cases for GR4J and TOPMO respectively.

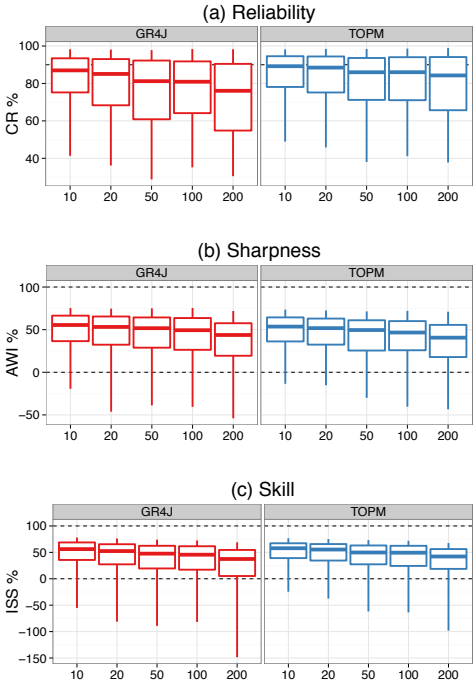


Figure 8. Impact of the hydrometrical desert on the distributions of the three performance criteria. Potential donors catchments are not retained as donors when their distance to the target catchment is below 10, 20, 50, 100 and 200 km. Boxplots (5th, 25th, 50th, 75th and 95th percentiles) synthesize the variety of scores over the 907 catchments of the data set.