# 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 understand from the comments that several parts of the manuscript need 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 intend to make in the manuscript to answer the major comments consist in:

- improving the description of the proposed approach;
- 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;
- 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.

Therefore we intend to resubmit a modified version of the manuscript. Specific changes planned for each review comments are explained below.

## 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.

> *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.

> *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.

> *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.

> *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.

> *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.

> *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.

> *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.

*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.

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

Thanks.

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

Thanks.

*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.

*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.

*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.

*P8050, L2: 'rainfall-runoff MODEL'*

Thanks.

*P8050, L26: 'increase' should be plural.*

Thanks.

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

Thanks.

> *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.

## 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.

> *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.

> *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.

> *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.

> *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.

> *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.

> *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.

> *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.

> *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.

> *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.

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

Agreed. This point needs a clearer presentation.

> *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.

## 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".

*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.

*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.

> *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.

> *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.

> *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.

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

Thanks. We mean hydrograph.

> *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.

> *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.

> *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.

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

Thanks.

*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.

*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.

*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.

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

Agreed.

*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.

*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.

*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).

*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.

*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.

## 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.

> *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.

> *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.

> *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.

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

Thanks.

> *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.

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

Thanks.

## 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".

> *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.

> *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.

> *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.

> *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.

> *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.

> *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.

> *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.