Reply to the review comments on the manuscript "A crash-testing framework for predictive uncertainty assessment when forecasting high flows in an extrapolation context" by Berthet *et al.* (manuscript hess-2019-181)

We first thank both referees for their detailed reviews and analyses of the submitted article. We also thank them for their positive opinion about the scientific soundness of this study and their constructive comments. They are very valuable to improve the manuscript and we intend to follow most of them (see details hereafter).

Both referees share some comments:

1. In the description of the methodology, the empirical hydrological uncertainty processor (EHUP) needs a better and more detailed description. Indeed, since there are some (quoted) references, we drastically reduced this description. It is clearly not sufficient and we propose to provide a more detailed description in a revised version.
2. The discussion deserves a better organization and some issues may be presented in the results section. We agree that this section needs to be reorganized. Indeed, we tried to build the article with a few 'seminal' "questions" motivating the study in the scope . Some additional questions appeared in the study and the discussion of the results. In order to improve the readability of the study, we will add some 'supplementary questions' in the scope, moving the corresponding results in the results section.
3. There are too many figures. In order to reduce the number of figures in the text, some will be removed, some will be merged and some will be moved to supplementary materials.

Below we give more detailed answers to the comments made by the reviewers and make some proposals explain how we propose to modify the text if the Editor request a revised submission.

Answer to the comments of the referee #1

General comments

This paper presents an approach for calibrating and evaluating extrapolated probabilistic hydrological predictions in the context of flood prediction. The authors consider a range of transformations for use in an uncertainty processor, and perform analysis over a large number of catchments, using multiple metrics to evaluate performance of the forecasts. The authors find that more complex transformations, which require calibration of parameters, may perform better over a calibration data-set, but typically do not perform best in an extrapolation context.

This is an interesting paper on an important topic, and is particularly relevant with a changing climate, where larger flooding events outside the range of historical observations may occur. The evaluation is comprehensive (large number of catchments, multiple metrics) and their analysis supports the key findings. However, I found that

(i) the description of the uncertainty processor, and in particular the role of the transformations, was insufficient, and

(ii) the discussion section requires additional work to explain the motivation for the additional analysis in this section.

Therefore, I recommend major revisions be made to this article before it can be published in HESS.

As explained in the general answer above, we agree with these general comments and changes will be made accordingly. See more details below.

> *Specific comments*
>
> *More details of EHUP. The empirical hydrological uncertainty processor (EHUP) and the different transformations are described in Section 2.1.3 and 2.1.4, respectively. Unfortunately, the level of detail provided by the authors was not sufficient to allow me to understand how the EHUP works and how the transformations fit inside the EHUP. In particular, It is unclear how the transformations fit in with EHUP. A diagram or mathematical equations would help make this clearer.*

EHUP deserves indeed a more detailed description. Since the variable transformation impacts the uncertainty assessment in an extrapolation context, the role of the variable transformation within the uncertainty processor will be presented in more detail.

> *It is not clear what are the inputs and outputs of the EHUP.*

We will clarify that the EHUP relies on the residuals of the discharge values available in the training data (inputs) and results in the conditional predictive distribution of the forecasted discharge.

> *Why does the EHUP require a separate training period to transformation calibration period?*

The EHUP is the non parametric method that 'only' needs a training period to "build" itself, i.e. to assess the empirical residual distributions on the different variable ranges. Moreover, some of the data transformations are parametric and require a calibration data set. In order to calibrate the transformation parameters, it is necessary first to produce these empirical residual distributions. We will clarify this point in the revised version.

> ***Discussion section.*** *The motivation for a lot of the analysis performed in this section is not clear to me, doesn't seem to fit in with the aims of the study, and often does not seem to match the sub- headings*
>
> *– Section 4.1*
>
> *o Pg 25, lines 9-17: This paragraph doesn't seem to be addressing the heading of the section, which is on the number of parameters in the transformation.*

We thank the referee for pointing out that the subsection title is not appropriate. We agree. The title will be changed. We will also rephrase the 2nd paragraph to explain the link with the 1st one.

> *– Section 4.2*
>
> *o This sub-section seems like it is attempting to determine what the key drivers for performance are, but this is not evident from heading.*

Indeed, this subsection title will be rephrased as a question to describe more explicitly the section content: "What are the possible drivers for performance losses when extrapolating?". Furthermore, this question will be added to the scope as a "supplementary question" and the section will be moved to the results section accordingly.

> *o Since the authors did not find any key drivers for poor performance, I'm not sure if this analysis adds much value.*

We agree with the referee on the fact that these negative results can be frustrating and do not bring much operational value. However, being able to explain when and how the performances decrease in an extrapolation context (*e.g.*, for very large and damaging floods) would be very valuable for

operational forecasters. Therefore, we think that it is important to mention that the possible 'drivers' we tested are not actual drivers. The motivation will be better explained. Furthermore, this subsection will be shortened (in particular, some figures removed or moved to the supplementary materials).

> *Section 4.3*
>
> *o The motivation for this section is unclear to me. Why are you comparing empirical and distribution based uncertainty? This seems tangential to the aims of the study. A brief sentence at the start to explain what you're looking into, and why, would be useful.*
>
> *o You are comparing "empirical-based" and "distribution-based" uncertainty assessment in this section. Since you have not explained the EHUP in enough detail, it is not clear which of these 2 approaches you have used for the rest of this study.*

We agree with the reviewer that the motivation for this section has to be better explained. Many studies are based on methodologies combining the use of data transformations and the assumption of a Gaussian distribution [Li et al, 2017]. Morawietz et al (2011) tested this issue specifically. This is will better explained in the revised text. Furthermore, the description of the link between the variable transformation and the characterisation of the distribution (EHUP) intended in section 2.1.3 (see above) will also contribute to make the motivation clearer.

> *– Section 4.4*
>
> *o This section is about making links to previous studies, but you cite only one paper and make no comparison to the findings of that paper.*

We agree with this remark. Since there are very few papers on this issue, we do not have enough materials to carry out a full comparison with previous studies. We will remove this subsection and add a few sentences on the link to McInerney et al. (2017) in the results section and/or the conclusion section.

> *– Section 5: "A need for a focus change."*
>
> *o This section is not long enough for a separate section, and is a discussion topic. I suggest moving this to the discussion.*

We thank the reviewer and will follow his/her suggestion.

> *– Limitations and future work*
>
> *o It would be useful to have a sub-section discussing the limitations of this study and future research.*

We agree with the fact that we need to better describe the limitations and the subsequent future research. A subsection or a specific paragraph will be dedicated to the limits and perspectives.

> *Too many figures. I believe this paper has too many figures. I recommend*
>
> *– Merging some figures*
>
> *o fig 6 and 10*
>
> *o fig 12 and 13*
>
> *– Is there any point in showing all 3 transformations for fig 16-18?*
>
> *o You could consider a single transformation and combine into a single figure.*
>
> *o Or you could move fig 17-18 to sup mat since they don't show any correlations.*

As mentioned in the general answer above, some figures will be merged or removed. Figures 12 and 13 will be merged, but we prefer to keep figures 6 and 10 separated because they are described in the text at two different places. We will move the figures 17 and 18 to the supplementary material.

> *Figure 5: I like the idea of having a diagram to explain how the different sets D1, D2sup, D2inf and D3 are used in calibration and evaluation, but I found this figure particularly confusing. In particular,*
>
> *– Why is different data used for EHUP in calibration and evaluation?*
>
> *– Why does D1 have many more points than D2sup and D3? From the text I thought D2sup and D3 had 720 points, while D1 had 500 points?*
>
> *– What's the purpose of showing the residuals on the y-axis? These are not discussed in the text.*
>
> *In panel b, most of the points for D2inf (light blue) are hidden behind points for D1 (red).*

We agree that more details are needed in the EHUP description. We will improve the description of the methodology up to subsection 2.1.4 and include more details to better understand the methods. The legend of figure 5 will be more detailed as well and we will clarify the selection of the 500 points for D1. The difference of the data used in the two steps will be better described in subsection 2.2.4.

The residuals are a key to understand the behaviour and the effects of the transformation. That is why they are discussed in section 4 (they are very important in the discussion in subsection 4.3). This will be mentioned in section 2.1.4.

> *Technical corrections*
>
> ***Abstract:** "... the Box-Cox transformation with a parameter between 0.1 and 0.3 can be a reasonable choice for flood forecasting"*
>
> *You have only shown results for lambda=0.2 in this paper. How you can say that using lambda between 0.1 and 0.3 can be a reasonable choice?*

We agree that this result is not shown in the submitted version: as explained in the methodology section, we studied a large number (17) of parameter values but we did not show the results for all of them, for the sake of brevity. The Box-Cox transformation has a "smooth" effect with respect to λ. We will add a figure in supplementary material.

> ***Table 1:** Change "percentiles" to "quantiles"*

This word will be changed.

> ***Pg 5, line 10:** "For each catchment, the lag time LT is estimated as the lag time maximising the cross-correlation between rainfall and discharge time series."*
>
> *What is the relevance of estimating LT? This becomes more obvious later in the paper, but should be described briefly here.*

Lag time is relevant to describe the catchment behaviour in a forecasting purpose: this characteristic duration has to be compared to the lead time (a) for the data assimilation procedures (most operational forecasting models use some) and (b) the relative importances of observed and forecasted precipitation inputs (which can explain part of the predictive performance when real precipitation forecasts are used). This will be better explained.

> ***Pg 5, line 15:*** *"It is a deterministic lumped storage-type model that uses catchment areal rainfall and PE as inputs"*
>
> *What rainfall is used to produce the GRP forecasts? Is observed rainfall used, forecast rainfall, etc? If it is observed rainfall, then how is this used in a forecasting context?*

We used the framework designed by Krzysztofowicz *et al.* in various studies, which separates the input uncertainty (mainly the observed and forecasted rainfall) and the hydrological uncertainty. This study focuses only on the 'effect' of the extrapolation degree in the hydrological uncertainty when using the best available rainfall product. In a forecasting context, when using uncertain rainfall, we will combine input uncertainty (rainfall) and hydrological uncertainty, as done for example in Bourgin et al. (2014).

> ***Pg 6, line 3-4:*** *"Since herein only the ability of the post-processor to extrapolate uncertainty quantification is studied, the model is calibrated in forecasting mode over the 10-year series by minimising the sum of squared errors for a lead time taken as the lag time LT."*
>
> *What is meant by "forecasting mode" here? More details on how forecasts are generated would be useful.*

The "forecasting mode" is to be compared to the "simulation mode" where no data assimilation is used. The latter allows to test the simulation model alone and assess its 'own' performance. The former is used to test a model in a context which is closer to the operational context (of the Flood Forecasting Service). Some references and a reference to appendix (where this is explained) will be added.

> ***Pg 7, lines 8-12:*** *Is the NQT actually used in this study? If so, it's not clear how and where it's used.*
>
> ***Pg 7, lines 14-15:*** *If the NQT requires additional assumptions for the tails, how do you handle this problem in in this study?*

We thank the referee for pointing out that this point is unclear. NQT was not tested, mainly because this transformation is known to require a particular care in an extrapolation context (see the technical note by Bogner *et al.* (2012) who explained that additional assumptions have to be made). However, since it is a frequently used transformation, we think that it is relevant in the introduction section. We will move this description at the very end of the subsection and explain why it was not used.

> ***Pg 7, lines 30-31:*** *"McInerney et al. (2017) obtained their best results with lambda = 0.2 over 17 perennial catchments."*
>
> *What do you mean by "best results"? Please provide some context for this statement.*

We used the paradigm set by Gneiting *et al.* (2007): the results are the "best" in terms of (1) reliability and (2) sharpness.

> ***Pg 8, line 3:*** *Why does this equation use different notation than other transformations?*

We thank the referee for pointing this inconsistency, which could be confusing for the reader. The notation will be made homogeneous.

> ***Pg 10, line 6:*** *"maximum discharge of time series"*
>
> *Make it clear you are referring to forecast discharge here.*

Changes will be made accordingly to this suggestion.

> ***Pg 10, line 8:*** *"the first time step"*
>
> *What do you mean by "first time step"? Do you mean the closest time step?*

We will better explain that the first time step of the event is the closest time step preceding the peak time step such as all discharge values from this time step to the peak are larger than 20 % (25%) of the peak flow value.

> ***Pg 10, lines 20-21, Pg 11, line 1:*** *The purpose of the "control", "training", and "calibration" subsets has not been explained. Please describe what they are used for.*

A short paragraph will be added to explain why the use of a variable transformation within an empirical HUP requires to use three subsets to test the performances in an extrapolation context.

> ***Pg 12, line 1:*** *It is unclear what the "coverage rate of the 80% predictive intervals" is. Please provide equations or description.*

We will add that these '80% predictive intervals' are bounded by the 0.1 and 0.9 quantiles of the predictive distributions.

> ***Pg 13, line 6:*** *"i.e. from the distribution of the observed discharges over the events selected"*
>
> *What is meant by "events selected"? Is this all events in G1, G2 and G3?*

We will clarify that the "events selected" refer to the events in the data subset for calibration or control.

> ***Pg 15, lines 6-8:*** *"as expected, and that there is no significant difference between the calibrated Box-Cox transformation (d), the calibrated log-sinh transformation (e) and the best performing transformation (f)."*
>
> *How are you determining whether differences between results are "significant"? A statistical test should be used to determine whether differences are "significant".*

A Mann-Whitney test has been used. It showed no significant difference between the reliability criterion values distributions obtained with the calibrated transformations. However, what we meant here is mainly that no difference can be noticed from Fig. 6. This paragraph will be rephrased in order to refer to what can be inferred from Fig. 6.

> ***Pg 15, line 8:*** *Is "best performing" the same as "best calibrated"? If so, use a single term.*

We thank the reviewer for pointing out that this difference could be somewhat confusing. A single expression will be used.

> ***Pg 15, lines 9-10:*** *"Interestingly, the log transformation provides the best results for the other criteria (not used as the objective function)."*
>
> *Are these results shown anywhere? If so, provide reference to figure.*

This sentence will be removed (see the answer to the second referee).

> ***Pg 15, line 13-14:*** *"While the log-transformation behaviour is frequently chosen for LT/2 and LT, the additive behaviour becomes more frequent for 2 LT and 3 LT."*
>
> *It is unclear what you mean by "additive behaviour" and how this is seen in the figures (i.e. what parameters relate to additive behaviour).*

The additive behaviour refers to the behaviour of the no transformation. A link to subsection 2.1.4. (page 7) where this is detailed, will be added.

> ***Pg 17, lines 7-8:*** *"This confirms that the CRPSS itself is not sufficient to evaluate the adequacy of uncertainty estimation"*
>
> *Similar findings about CRPS being insensitive to chosen data transformation have been made in other studies, e.g. Woldemeskel et al. (2018). It might be worth mentioning this.*

We did not know this article. Thank you for giving this reference. We will mention it in the revised article.

> ***Figure 9 caption:*** *"Thérain River at Beauvais (755 km2): the forecasts are reliable and ...."*
>
> *This statement does not seem correct. I would say the forecasts are not reliable for "none".*

This comment is very true, we implicitly described only the uncertainty assessment when a variable transformation is used, since such a transformation is most often needed to achieve (more or less) reliable results. "(except if no transformation is used)" will be added.

> ***Figure 14:*** *Legend is missing*

We apologize for this missing legend. Legend is given below the figure.

> ***Pg 25, line 22-23:*** *"The results indicate that it is not possible to anticipate the alpha-index values when extrapolating high flows in D3 based on the alpha-index values obtained when extrapolating high flows in D2sup."*
>
> *There appears to be some correlation in Figure 16. What is the Spearmen correlation?*
>
> ***Pg 25, lines 27-28:*** *"In both cases, no trend appears, regardless of the variable transformation used, with Spearman correlation coefficients lower than 0.5."*
>
> *A Spearman correlation of 0.5 does not seem correct. If it was 0.5, then there would be a clear trend.*

Spearman values were all lower than 0.33 .

> ***Pg 25, line 31:*** *What is a "normalized RMSE" and why is it used? A sentence/equation describing this would be useful (rather than just a citation).*

We thank the referee for having detected the absence of description of this criterion. A description will be added in section 2.3.1.

> ***Pg 26, line 11:*** *"Even if there is no theoretical advantage to using the Gaussian distribution calibrated on the transformed-variable residuals rather than the empirical distribution to assess the predictive uncertainty, we tested the impact of this choice."*
>
> *If there is no theoretical advantage, why are you testing this?*

As said in the general answer, the motivation of the subsection 4.3 will be better explained at its beginning.

> ***Pg 33, line 21:*** *"the Box-Cox transformation with its lambda parameter set at 0.2 or between 0.1 and 0.3."*
>
> *You have only shown results for lambda=0.2 in this paper. How you can recommend using other values of lambda between 0.1 and 0.3 in the conclusions of this paper?*

As explained above, this result was indeed not shown in the submitted version for the sake of brevity but we studied a large number (17) of parameter values. We will add a figure in the supplementary material.

> ***Section B2.2.*** *This relationship was shown by McInerney et al. (2017) (Appendix A)*

We agree that this relationship was also pointed out by McInerney *et al.* (2017). This will be acknowledged in the text.

> ***References***
>
> *McInerney, D., Thyer, M., Kavetski, D., Lerat, J. & Kuczera, G. 2017. Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. Water Resources Research, 53.*
>
> *Woldemeskel, F., McInerney, D., Lerat, J., Thyer, M., Kavetski, D., Shin, D., Tuteja, N. & Kuczera, G.2018. Evaluating residual error approaches for post-processing monthly and seasonal streamflow forecasts. Hydrol. Earth Syst. Sci. Discuss., 2018, 1-40.*

Answer to the comments of Dr. Engeland (referee #2)

> *The paper presents a framework aiming at evaluating the performance of probabilistic forecasts on highest flood events that the post-processors are not calibrated for. The authors combine an empirical hydrological post-processor (EHUP) with different trans-formations, and compare the performance of the predictive distributions for forecasted floods that are higher than the floods used for training/calibrating the EHUP and the transformations.*
>
> *The paper is interesting and deserves publication following a major revision. Below It lists some important issues to be addressed in the revised manuscript.*
>
> *Throughout the introduction, the importance of modelling the heteroscedasticity of the predictive uncertainty distribution emphasized. I miss a good argument why it is important (i.e. to obtain reliability), and you could refer to literature that shows this (i.e. McInerney et al., 2017). In the introduction and discussion, you ignore that other properties of the predictive distribution (i.e. bias and skewness) could also depend on forecasted flows. My experience is that a calibrated hydrological model tends to under-estimate flood peaks, introducing a possible bias. Bremnes et al (2019) shows that the skewness depends on the predicted wind and that adding this property improves the forecasts for high wind speeds. You discuss this briefly in lines 4-9 on page 8. Is it possible that the results presented in Figures 12 and 13 indicate that the skewness is an important issue for the reliability of the predictive distributions, and that your approach has to small skewness?*

The referee points out here an important fact. The heteroscedasticity is an important property to describe a probability distribution and is often looked out in the literature, but it is very true that all the properties of the distribution have to be checked. This issue will be mentioned in the introduction and the conclusion.

We are not sure that figures 12 & 13 give any indication on the skewness. They only describe the reliability of two predictive quantiles. They show that the evolution in an extrapolation context of the empirical distribution assessed by EHUP is not perfectly reliable.

Indeed, the main issue here is the stability of the overall predictive distribution in an extrapolation context. The tests provided in section 4.3 give some insights.

I miss an explanation of which meteorological products you used to generate the discharge forecasts. Did you use the reanalysis mentioned in 2.1.1 or did you use a forecast product?

We used the reanalysis mentioned in 2.1.1 as meteorological inputs. We chose to follow the decomposition proposed by Krzysztofowicz (input uncertainty and modelling uncertainty): here we test only the modelling uncertainty in extrapolation. Further work shall investigate the contribution of the input uncertainty (Bourgin, 2014) in an extrapolation context. This will be mentioned in section 2.1.1 and in the conclusion.

The EHUP needs a better description, in particular how it is used in combination with the different transformations. I also need a clarification of which data were used for estimating the empirical quantiles of errors. On page 6 you write that the top 5% pairs ranked by simulated values are used, whereas on page 11 you write that the subsets D1 and D2 were used. Figure 5 indicates that not the whole D1 subset was used for training of the EHUP, only the highest discharge values. A consistent description is needed to avoid confusion.

We thank both reviewers and agree with them on the fact that EHUP needs a better description. It will be done following their comments. Note that the 5%-selection is made on the subset used for the training (D1 for the calibration step and D1 + D2 for the evaluation step). Figure 5 and its legend will be improved to make clear that only the top 5% pairs are used for the extrapolation.

The discussion section needs a better organization. Results presented in section 4.1could be integrated into section 3. In Figure 15, the only new result is the boxes labelled 'g'. Could it be integrated into Figure 10? Section 4.2 and 4.3 introduces new results that do not directly relate to the objectives / questions listed on Page 4. If these results should be included, you could add one more objective related to these results, and integrate the results into Section 3. I suggest to exclude results and discussion in section 4.3 (including Figure 19 and 20) and only briefly summarize the main findings.

As mentioned in the general answer above, we agree and look forward a better organization of the section. We prefer to keep subsections 4.1 and 4.3, because they mostly bring information to interpret the main results. In order to do so, we achieved a few complementary tests. The issue in subsection 4.3 seems particularly important because this assumption is often made but sometimes not tested. The scope (subsection 1.3) will be completed in order to make it appear at the beginning of the article. We will place the 2nd figure in the supplementary materials.

Section 5 could be also a part of the discussion.

We agree. Section 5 will be included as the last subsection in the discussion.

The number of figures could be reduced. Figure 2 – right panel is not necessary.

We agree that both panels of figure 2 are not necessary, but we prefer keeping the right panel because it better explains the effect of the transformation on the uncertainty assessment: a constant probability distribution in the transformed space will evolve in the untransformed space based on the behaviour of the inverse data transformation.

Figures 4a and 4b could be merged. Is it possible the merge Figure 5a and b? Could result sin Figure 15 be included in Figure 10? Figure 11 is not necessary. Figure 12, and 13could be merged. I suggest to remove Figure 14a since it is just another measure of reliability and does not add new information to the results. Figures 19 and 20 could be excluded or moved to supplementary materials.

We thank the referee for pointing out that some figures can be rearranged or merged. Figures 4a and 4b will be merged. However we did not manage to merge figures 5a and 5b in a unique meaningful and easy-to-read figure. Results in figure 15 will be added to figure 10. We respectfully disagree on figure 11, which we consider interesting since it is the only one displaying a scatter plot (while most of the figures display box-plots), which brings an additional and valuable information: the comparison catchment per catchment. Figures 12 and 13 will be merged. Figure 14a provides indeed another reliability criterion but this one brings another information (both α-index and coverage ratio criteria are synthetic criteria) and is important for many operational forecasters.

> *Below follows some detailed comments to the manuscript:*
>
> *Table 2: When you compare discharge across catchments, I think it is better to use specific discharge (l/s/km²).*

We agree. Peak discharges will be describe through specific discharge values.

> *Figure 3: What is the explanations for this apparently negative skewness for the predictive distribution? The log-transformation leads to slightly positively skewed predictive distribution?*

The empirical distribution provided by EHUP reflects the assessed distribution on the training data set. The log transformation exacerbates the skewness, since it has a “multiplicative effect”.

> *Figure 14: Legend is missing*

We apologize for the missing legend. Legend is given below the figure.

> *Page 2: The meaning of the first paragraph of section 1.2 is difficult to understand. In particular the two first sentences needs more context.*

We thank the referee for this warning. The paragraph will be rewritten, giving the context of operational forecast systems and organization, in order to provide useful information to crisis managers.

> *Page 3: I suggest to write the first paragraph of 1.2.1 as: “A first approach that intends to model each source of uncertainty separately and to propagate these uncertainties through the modelling chain is presented in Renard et al., (2010). According to this approach, the heteroscedasticity of the predictive uncertainty distribution results from the separate modelling of each source of uncertainty and from the statistical model specification. While this approach is promising, operational application can be hindered by the challenge of making the hydrological modelling uncertainty explicit, as pointed out by Salamon and Feyen (2009).”*

We thank the referee and adopt his proposal.

> *Question or the paragraph above: which statistical model needs to be specified? Is it for the meteorological input or for the simulated discharge?*

Renard et al. (2010) use a Bayesian modelling, which needs a full specification of the inputs distribution (assumptions) and of the likelihood (another assumption).

> *Page 4 lines 7-8: These approaches are not exclusive of each other. Even when future precipitation is the main source of uncertainty, post processing is often required to produce reliable hydrological ensembles Question: What does ‘these approaches' refer to? does it refer to all approaches presented in the introduction or all approaches presented in section 1.2.2?*

We agree that this sentence is not clear. “These approaches” refer to the two main families described in subsections 1.2.1 and 1.2.2. This will be specified in the revised manuscript.

> *Page 5 Section 2.1.1: Maybe a question of style, you write ‘We used a set of 154 unregulated catchments spread throughout France (Fig. 1) to test our hypotheses over various hydrological regimes and forecasting contexts." Since you have chosen to use formulate research questions and not to test hypotheses in this paper, the sentence could be changed to ‘We used a set of 154 unregulated catchments spread throughout France (Fig. 1) over various hydrological regimes and forecasting contexts to provide robust answers to our research questions.*

We agree and will change the text accordingly.

> *Page 7, line 19: You write that the log-transformation is non-parametric. I would rather say it is a parametric transformation with no tuning parameters. The term non-parametric is often used when you make no assumptions about the form or parameters of the transformation.*

We agree that the term “non-parametric” is frequently used for distributions and means that there is no assumption about the form of the distribution. This word can also be used for transformations of functions. Then it only refers to the existence of tuned parameters. We will clarify the meaning in the text.

> *Page 10 Section 2.2.2. How did you select more than one event? According to the description you selected one event defined by the maximum discharge of the time series.*

Once the first event is selected, the process is iterated over the remaining data to select more events. This point will be detailed in the revised text.

> *Page 11: Why has the calibration data subset to encompass time steps with simulated discharge values higher than those of the training subset?*

Since our intention is to test the robustness and adequacy of different data transformations in an extrapolation context, it is more useful to calibrate their parameters in an extrapolation context, i.e. on simulated discharge values larger than the ones met in the training step. In addition, since we used an empirical uncertainty processor, the data transformations have almost no impact on the uncertainty estimation in the training subset and we will not be able to “tune” their parameters.

> *Page 13: First equation: define k and N Second equation: Could you use the same notation as in the first equation. i.e. write it as sum divided by number of time steps?*

N is the number of time steps on which the CRPS is computed and k is just an index. We will precise the meaning of N and write the second equation using the same notation.

> *Page 15, lines 9-10: Here you comment results that are not yet presented, making it difficult for the reader to follow. I think this sentence fits better in the discussion*

We thank the referee for his careful review. This sentence corresponds to some results that were not shown. It will be removed in the revised manuscript.

> *Page 16: The last three lines have to be re-phrased in order to make sense: "In operational settings, non-exceedance frequencies of the lower (0.1 quantile) and the exceedance frequencies of the upper (0.9 quantile) bounds of the predictive distribution are of particular interest. It is expected that those values remain close to 10% for a reliable predictive distribution. Deviations from these frequencies indicates biases in the estimated quantiles."*

We thank the referee for his proposal. The sentences will be rewritten.

> *Page 17 lines 3-5: I think it is better to write something like this (I think it is better to write that the 0.1 and 0.9 quantiles are over or under-estimated, and not the (non)-exceedance frequency of the (0.1) and 0.9 quantiles.): "More importantly, it can be seen that the lack of reliability of the log transformation seen for 3 LT in Fig.10 appears to be related to an underestimation of both the 0.1 and 0.9 quantile. Compared to the other transformations, the log transformation has the largest under-estimation of the0.1 quantile and the smallest under-estimation of the 0.9 quantile."*

The sentences will be rewritten to make them clearer.

> *Page 18 Section 3.2.2: Please be more precise in the comments: What is 'overall performance' ? Suggestion for re-phrasing some of the sentences:"We note that the log transformation has the highest median value for the coverage ratio, and is also the closest to the 80% ratio that is expected from a reliable forecast,""In addition, the CRPSS and the NSE distributions have limited sensitivity to the variable transformation. We can even see that not using any transformation yields slightly better results according to NSE.*

The "overall performance" refers to an "overall" criterion which does not investigate a specific property of the forecasts (reliability, accuracy, sharpness…) but intends to describe the whole predictive distribution. We used the CRPS, as mentioned in subsection 2.3.1. It will be written in section 3.2.2 as well to make it clearer.

> *Page 33: Please provide clear conclusions related to each of the objectives and answer the research questions asked in the introduction.*

We thank the referee for this suggestion that we will follow in the revised version of this article.

> *New reference in this review: Bremnes, J.B., 2019: Constrained Quantile Regression Splines for Ensemble Post processing. Mon. Wea. Rev., 147, 1769–1780,https://doi.org/10.1175/MWR-D-18-0420.1*

References

Bogner, K., Pappenberger, F. and Cloke, H. L. (2012). Technical Note: The normal quantile transformation and its application in a flood forecasting system *Hydrology and Earth System Sciences*, **16**: 1085-1094

Li, M., Wang, Q. J., Robertson D. E. and Bennett J. C. (2017). Improved error modelling for streamflow forecasting at hourly time steps by splitting hydrographs into rising and falling limbs. *Journal of Hydrology*, **555**: 586-599. https://doi.org/10.1016/j.jhydrol.2017.10.057

Morawietz, M., Xu, C.-Y., Gottschalk, L. and Tallaksen, L. M. (2011). Systematic evaluation of autoregressive error models as post-processors for a probabilistic streamflow forecast system *Journal of Hydrology*, **407**: 58-72