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

Line in the answer to referees' comments (*)	Page in the first submitted version	Line in the first submitted version	Changes made in the second submitted version
328	1	8	The sentence has been completed as: "*to account for heteroscedasticity and the evolution of the other properties of the predictive distribution with the discharge magnitude.*"
402	2	21	The beginning of the subsection 1.2 has been rewritten to better refer to the context of operational services: *"Even if significant progress has been made and implemented in operational flood forecasting systems (e.g., Bennett et al., 2014; Demargne et al., 2014; Pagano et al., 2014), some uncertainty remains. In order to achieve an efficient crisis management and decision making, communication of reliable predictive uncertainty information is therefore a prerequisite (Todini, 2004; Pappenberger and Beven, 2006; Demeritt et al., 2007; Verkade and Werner, 2011). Hereafter, reliability is defined as […]".*
328	2	30/31	The sentence has been rewritten as: "*In an extrapolation context, it is of utter importance that the predictive uncertainty assessment provides a correction description of the evolution of the predictive distribution properties with the discharge magnitude. Bremnes (2019) showed that the skewness of wind speed distribution depends on the forecasted wind. Modelled residuals of discharge forecasts often exhibit high heteroscedasticy (Yang* et al., *2007). McInerney et al. (2017) focused their study on representing error heteroscedasticity of discharge forecasts with respect to simulated streamflow. To achieve reliable forecasts, a correct description of the heteroscedasticity, either explicitly or implicitly, is necessary.*
413	3	6	The first paragraph of the subsection1.2.1 has been rewritten 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).*"
328	4	13	The sentence has been rewritten as: "*Note that many of these approaches use a variable transformation to handle the heteroscedasticity (and more generally the evolution of the predictive distribution properties with respect to the forecasted discharge).*"

Line in the answer to referees' comments (*)	Page in the first submitted version	Line in the first submitted version	Changes made in the second submitted version
	4		The last sentence of the first paragraph of the Scope subsection has been rewritten as: "*Since achieving a reliable predictive uncertainty assessment in an extrapolation context is a challenging task likely to remain imperfect if the stability of the characteristics of the predictive distributions is not properly ensured, it requires a specific crash-testing framework~\citep{Andreassian2009}. The objectives of this article are:*"
79 & 363	4	28	The "Scope" subsection has been completed with a 3rd question: « *We attempt to answer three questions : (a) Can we improve […] ? (b) Do more flexible transformations […] ? (c) If there is a performance loss when extrapolating, is there any driver that can help the operational forecasters to predict this performance loss and question the quality of the forecasts ?* »
423	4	###	A new subsection has been set up to more clearly 'separate' the two last paragraphs and the sentence has been rewritten as : "*The approaches presented in subsections 1.2.1 and 1.2.2 are not exclusive of each other*."
432	5	4	The sentence has been rewritten as: "*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.*"
166	5	10	A sentence has been added to point out the importance of LT and to test different lead times: "*When a hydrological model is used to issue forecasts, it is often necessary to compare the lead-time to a characteristic time of the catchment (section 2.1.2). For each catchment [...]*"
187	5	15	The last sentence has been completed as: "*In forecasting mode (appendix A), the model also assimilates [...]*"
161	5		The word "*Percentiles*" have been replaced by the word "*Quantiles*" in Tab. 1
166	6	6	A reference to Berthet *et al.* (2009) has been added.
46 & 54 & 352 54	6	7	The description of the EHUP (mainly in subsection 2.1.3.) has been modified in deep, so that it brings information enough to make the methodology clear. The role of the residuals and of the inverse transformation have been emphasized.

(*) In *190803_Answers to revisions.pdf*. Before line 307 (blue): 1st (anonymous) referee ; after line 30 (green) : 2nd referee (Dr. Engeland)

Line in the answer to referees' comments (*)	Page in the first submitted version	Line in the first submitted version	Changes made in the second submitted version
143 & 352			The meanings of the selection of the 500 top values in D1 and the role of the 5%-highest values have been clarified. The legend of Fig. 5 has been rewritten to make clearer that only the information conveyed by the top 5% pairs of the training data subset are used by EHUP to extrapolate on the calibration or control periods.
175 & 340	6	4/5	A sentence is added after the sentence whose 1st words are "*Since herein only the ability of the post-processor to extrapolate uncertainty quantification is studied*": "*For the same reason, the model is fed only with observed rainfall (no forecast of precipitation), in order to reduce the impact of the input uncertainty.*"
196	7	8	The NQT description has been sent at the end of the subsection 2.1.4 and starts with: "*Another common variable transformation is the normal quantile transformation (NQT). It is a [...]*" The following sentence has been added at the end of the paragraph: "*This is why we did not test this transformation in this study focused on the extrapolation context.*"
196	7	16	The sentence has been changed as: « *Three analytical transformations are often met in hydrological studies: [...].* »
437	7	19	The following words have been added at the end of the sentence "*This transformation is then non-parametric*": "*(no parameter has to be calibrated).*"
205	7	31	The sentence has been changed as: « *McInerney* et al. *(2017) obtained their most reliable and sharpest results with [...]* »
208	8	3	The equations formalism has been made homogeneous.

(*) In *190803_Answers to revisions.pdf*. Before line 307 (blue): 1st (anonymous) referee ; after line 30 (green) : 2nd referee (Dr. Engeland)

Line in the answer to referees' comments (*)	Page in the first submitted version	Line in the first submitted version	Changes made in the second submitted version
372	8		The left panel of the figure 2 has been removed: only the right panel is kept. A sentence has been added to its legend: "*the inverse transformations are displayed because they explain the final effect of the transformation on the uncertainty assessment: the constant probability distribution in the transformed space (provided by the EHUP) will result in an distribution in the untransformed space, whose evolution depends on the behaviour of the inverse data transformation.*"
212	10	6	The fact the selection is made on the forecasted discharge series has been clarified: « *(1) the maximum forecasted discharge of the time series was selected [...]* »
215	10	7	The words "*the first time step*" have been replaced by "*the preceding (following) time step closest to the peak*"
444	10	10	The following sentence has been added before the sentence which starts as "*A minimum time lapse of 24 h was enforced between two events […]*": "*The process is then iterated over the remaining data to select all events.*"
392	10	Tab. 2	The median value of the peak specific discharges are given (instead of peak discharges)
143	11	22	The difference between the data used for the training in the calibration step and in the evaluation step has been clarified in order to better explain the objective of the methodology: « *In the second step, the EHUP was trained on a data set which encompassed D1, D2inf and D2sup using the parameter set obtained during the calibration step. Then, the predictive uncertainty distribution was* *evaluated on the control data set D3. Training the EHUP on the union of D1, D2inf and D2sup allows to control the uncertainty assessment from small to large degrees of extrapolation (on D3). Indeed if we had kept the training on D1 only, we would have not been able to test small degrees of extrapolation on independent data for every catchment (see the discussion in Sect. 3.3).* »
448	11	23	The following sentences have been added before the sentence "*The parameter set obtaining the best criterion value was selected.*": "*Indeed, the data transformations have almost no impact on the uncertainty estimation by EHUP on events of the same magnitude as those of the training subset. Therefore the calibration subset has to encompass events of a larger magnitude (D2Sup).*"

(*) In *190803_Answers to revisions.pdf*. Before line 307 (blue): 1st (anonymous) referee ; after line 30 (green) : 2nd referee (Dr. Engeland)

Line in the answer to referees' comments (*)	Page in the first submitted version	Line in the first submitted version	Changes made in the second submitted version
224	12	1	The sentence has been completed as: « *[…] the coverage rate of the 80 % predictive intervals (bounded by the 0.1 and 0.9 quantiles of the predictive distributions […]* »
381	12		Fig. 4a and 4b have been merged. Furthermore, the legend has been modified: "*Illustration of the selection of the data subsets for the Ill River at Didenheim (668 km2). First, the events are selected (grey highlighting). Then, the four data subsets are populated according to the thresholds (horizontal dashed lines). See Sect. 2.1.3 for more details.*"
455	13	4	The sentence has been modified as "*where N is the number of time steps, F the predictive cumulative distribution, H the Heaviside function and […]*"
229	13	6	The words "*over the events selected*" have been replaced by "*over the same data subset*".
455	13	9	The second equation has been rewritten using the same notation. $\frac{\sum_{k=1}^{N} q_{0.9}(Q_k) - q_{0.1}(Q_k)}{\sum_{k=1}^{N} Q_{k,obs}}$ "*Where $q_{0.9}(Q_k)$ is the quantile 0.9 of the predictive distribution at time step k.*"
143	14		Figure 5. has been modified to better emphasize the role of the top 5%-forecasted discharge values.
236	15	7	The words "*there is no significant difference*" have been replaced by "*no noticeable difference can be seen in Fig. 6 between […].*"
242	15	8	The words "*the best performing transformation (f)*" have been replaced by "*the best calibrated transformation (f)*".
247 & 459	15	9	The sentence "*Interestingly, the log transformation provides the best results for the other criteria (not used as the objective function.*" has been removed.
252	15	14	The words "(corresponding to the use of no transformation, see *Sect. 2.1.4)*" have been added after "*additive behaviour*".

(*) In *190803_Answers to revisions.pdf*. Before line 307 (blue): 1st (anonymous) referee ; after line 30 (green) : 2nd referee (Dr. Engeland)

Line in the answer to referees' comments (*)	Page in the first submitted version	Line in the first submitted version	Changes made in the second submitted version
467	16	17/19	The sentence starting as "*In operational settings, non-exceedance […]*" has been rewritten as: "*In operational settings, non-exceedance frequencies of the quantiles of the predictive distribution which are the lower and upper bounds of the predictive interval communicated to the authorities are of particular interest. The 80%-predictive interval (bounded by the 0.1 and 0.9 quantiles) is often used. It is expected that the non-exceedance frequencies of the lower bound and the exceedance frequencies of the upper bound remain close to 10% for a reliable predictive distribution. Deviations from these frequencies indicates biases in the estimated quantiles."*
475	17	###	The sentence has been rewritten as: "*More importantly, it can be seen that the lack of reliability of the log transformation for the 3-LT lead time seen in Fig. 10 appears to be related to an underestimation of the 0.1 quantile which is more important than for the other tested transformations, while the 0.9 quantile is less underestimated than for the other transformations.*"
483	18	2	The words "*(measured by the CRPSS)"* have been added after the words "*namely the overall performance*"
259	18	6	The sentence is completed and becomes « *In addition, the CRPSS and the NSE distributions have limited sensitivity to the variable transformations (also shown by Woldemeskel* et al.*, 2018, for the CRPS), even if […]* »
265	19		In the legend of Fig. 9, the words "*(except if no transformation is used)*" are added at the end of the comment for the Thérain River at Beauvais.
381	20		Fig. 10 & 15 have been merged.
269 & 399	24		Legend of Fig. 14 is given below the figure.
74	25	2	The title of the 1st subsection of the discussion has been changed as: "*Do more complex parametric transformations yield better results in an extrapolation context ?*"
74	25	9	The first sentence of the 2nd paragraph of subsection 4.1 has been rewritten as: "*These results could be explained by the fact that the calibration did not result in the optimally relevant parameter set. To investigate whether […]*"

(*) In *190803_Answers to revisions.pdf*. Before line 307 (blue): 1st (anonymous) referee ; after line 30 (green) : 2nd referee (Dr. Engeland)

Line in the answer to referees' comments (*)	Page in the first submitted version	Line in the first submitted version	Changes made in the second submitted version
79	25	18	The title of the 2nd subsection of the discussion has been changed as: "I*nvestigating the performance loss in an extrapolation context*"
79 & 363	25	18	The subsection 4.2 has been moved to the results section (and becomes the subsection 3.3).
85	25	18	A sentence has been added to start the subsection dedicated ot the performance loss in an extrapolation context (initially subsection 4.2, now subsection 3.3) "*It is very important that operational forecasters can predict when they can trust in the forecasts issued by their models and when their quality becomes questionable. Therefore we investigated […]"*
278	25	27/28	The end of the sentence has been modified in : "*with Spearman coefficients values (much) lower than 0.33*"
100	26	1	The title of the 3rd subsection of the discussion has been changed as: "*Empirical-based versus distribution-based approaches : does the distribution shape choice impact the uncertainty assessment in an extrapolation context?"*
100 & 287	26	2	In order to better explain the objectives and motivations of this analysis, the subsection now starts by these sentences: « *Besides the reduction of heteroscedasticity, many studies use post-processors which are explicitly based on the assumption of a Gaussian distribution and use data transformations to fulfil this hypothesis [Li et al, 2017]. Some post-processors are based on it, such as the MCP or the meta-Gaussian model, and the NQT was designed to precisely achieve it. Morawietz et al (2011) tested this issue. We first checked whether the variable transformation helped to reach a Gaussian distributed of the residuals computed with the transformed variables. Then we investigate whether better performance can be achieved using eimpirical transformed residuals distributions or using Gaussian distributions calibrated on these empirical distributions.* *We used the Shapiro-Francia test. [...]* »
109	29	1	The subsection 4.4 has been removed. The links with the research by McInerney *et al.* (2017) have been recalled in the results and conclusion sections.
363	30		Fig. 20 has been placed in Supplementary materials

(*) In *190803_Answers to revisions.pdf*. Before line 307 (blue): 1st (anonymous) referee ; after line 30 (green) : 2nd referee (Dr. Engeland)

Line in the answer to referees' comments (*)	Page in the first submitted version	Line in the first submitted version	Changes made in the second submitted version
116 & 370	32	1	The section 5 has been included as the last subsection of the discussion.
328	33	9/10	The sentence has been rewritten as: "*The latter has to handle the heteroscedasticity and the evolution of the other predictive uncertainty distribution properties, which is very problematic in an extrapolation context to issue reliable uncertainty assessment.*"
489	33	12/…	The conclusion has been reorganized. It now includes a "*Main findings*" subsection and a "*Limitations and perspectives*" subsection. In the former, an answer is explicitly provided to every research question specified in the scope subsection (a), (b) and (c).
293	33	21	*A figure is provided in supplementary materials to show it.*
85 & 120 & 340	33	25	The subsection "*Limitations and perspectives*" of the conclusion includes: " *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. Future works should combine both input uncertainty (rainfall) and hydrological uncertainty (Bourgin et al., 2014), to evaluate the impact of using uncertain (forecasted) rainfall in a forecasting context.* *We found no variable correlated to the performance loss we observed in an extrapolation context. Testing more variables potentially correlated is necessary. First it may open new perspectives to explain these losses and improve our understanding of the flaws of the hydrological model and of the EHUP. Furthermore, it would be very useful to help the operational forecasters to detect major events when their forecasts have to be particularly questioned.* *Furthermore, improving the regionalisation of the predictive distribution assessment, as proposed in Bourgin et al. (2015) and Bock et al. (2018) could help build more robust assessment of uncertainty quantification when forecasting high flows.*"
297	37	13	The sentence has been changed in "*As pointed out by McInerney* et al. *(2017), when alpha << y […]*"
125 & 381	22 & 23		Fig. 12 & 13 have been merged

(*) In *190803_Answers to revisions.pdf*. Before line 307 (blue): 1st (anonymous) referee ; after line 30 (green) : 2nd referee (Dr. Engeland)

Line in the answer to referees' comments (*)	Page in the first submitted version	Line in the first submitted version	Changes made in the second submitted version
128	28 & 29		Fig. 17 & 18 have been moved to the supplementary materials
–	End		A sentence was added to acknowledge the contributions of Dr. Engeland and of an anonymous referee.

(*) In *190803_Answers to revisions.pdf*. Before line 307 (blue): 1st (anonymous) referee ; after line 30 (green) : 2nd referee (Dr. Engeland)

A crash-testing framework for predictive uncertainty assessment when forecasting high flows in an extrapolation context

Lionel Berthet[1], François Bourgin[2], Charles Perrin[3], Julie Viatgé[3], Renaud Marty[1], and Olivier Piotte[4]

[1]DREAL Centre-Val de Loire, Loire Cher & Indre Flood Forecasting Service, Orléans, France
[2]IFSTTAR, GERS, EE, Bouguenais, France
[3]IRSTEA, HYCAR Research Unit, Antony, France
[4]Ministry for the Ecological and Inclusive Transition, SCHAPI, Toulouse, France

Correspondence: lionel.berthet@developpement-durable.gouv.fr

Abstract. An increasing number of flood forecasting services assess and communicate the uncertainty associated with their forecasts. While obtaining reliable forecasts is a key issue, it is a challenging task, especially when forecasting high flows in an extrapolation context, i.e. when the event magnitude is larger than what was observed before. In this study, we present a crash-testing framework that evaluates the quality of hydrological forecasts in an extrapolation context. The experiment setup is based on i) a large set of catchments in France, ii) the GRP rainfall-runoff model designed for flood forecasting and used by the French operational services and iii) an empirical hydrologic uncertainty processor designed to estimate conditional predictive uncertainty from the hydrological model residuals. The variants of the uncertainty processor used in this study differ in the data transformation they ~~used~~ use (log, Box-Cox and log-sinh) to account for heteroscedasticity and the evolution of the other properties of the predictive distribution with the discharge magnitude. Different data subsets were selected based on a preliminary event selection. Various aspects of the probabilistic performance of the variants of the hydrologic uncertainty processor, reliability, sharpness and overall quality, were evaluated. Overall, the results highlight the challenge of uncertainty quantification when forecasting high flows. They show a significant drop in reliability when forecasting high flows in an extrapolation context and considerable variability among catchments and across lead times. The increase in statistical treatment complexity did not result in significant improvement, which suggests that a parsimonious and easily understandable data transformation such as the log transformation or the Box-Cox transformation with a parameter between 0.1 and 0.3 can be a reasonable choice for flood forecasting.

1 Introduction

1.1 The big one: dream or nightmare for the forecaster?

In many countries, operational flood forecasting services (FFS) issue forecasts routinely throughout the year and during rare or critical events. End-users are mostly concerned by the largest and most damaging floods, when critical decisions have to be made. For such events, operational flood forecasters must get prepared to deal with extrapolation, i.e. to work on events of a magnitude that they and their models have seldom or never met before.

The relevance of simulation models and their calibration in evolving conditions, such as contrasted climate conditions and climate change has been studied by several authors. For example, Wilby (2005), Vaze et al. (2010), Merz et al. (2011) and Brigode et al. (2013) explored the transferability of hydrological model parameters from one period to another and assessed the uncertainty associated with this ~~parametrization~~parameterization, while Coron et al. (2012) proposed a generalisation of the differential split-sample test (Klemeš, 1986). In spite of its importance in operational contexts, only a few studies have addressed the extrapolation issue for flow forecasting, to the best of our knowledge, with the notable exception of data-driven approaches (e.g., Todini, 2007). Imrie et al. (2000), Cigizoglu (2003) and Giustolisi and Laucelli (2005) evaluated the ability of trained artificial neural networks (ANNs) to extrapolate beyond the calibration data and showed that ANNs used for hydrological modelling may have poor generalisation properties. Singh et al. (2013) studied the impact of extrapolation on hydrological prediction with a conceptual model, and Barbetta et al. (2017) expressed concerns for the extrapolation context defined as floods of a magnitude not encountered during the calibration phase.

Addressing the extrapolation issue involves a number of methodological difficulties. Some data issues are specific to the data used for hydrological modelling, such as the rating curve reliability (Lang et al., 2010). Other well-known issues are related to the calibration process: are the parameters, which are calibrated on a limited set of data, representative or at least somewhat adapted to other contexts? A robust modelling approach for operational flood forecasting, i.e. a method able to provide relevant forecasts in conditions not met during the calibration phase, requires paying special attention to the behaviour of hydrological models and the assessment of predictive uncertainty in an extrapolation context.

1.2 Obtaining reliable forecasts remains a challenging task

Even if significant progress has been made and implemented in operational flood forecasting systems (e.g., Bennett et al., 2014; Demargne et al., 2014; Pagano et al., 2014), some uncertainty remains. ~~Communication~~ In order to achieve an efficient crisis management and decision making, communication of reliable predictive uncertainty information is therefore ~~required to improve crisis management and decision making~~a prerequisite (Todini, 2004; Pappenberger and Beven, 2006; Demeritt et al., 2007; Verkade and Werner, 2011). Hereafter, reliability is defined as the statistical consistency between the observations and the predictive distributions (Gneiting et al., 2007).

The uncertainty associated with operational forecasts is most often described by a predictive uncertainty distribution. Assessing a reliable predictive uncertainty distribution is challenging because hydrological forecasts yield residuals that show heteroscedasticity, i.e. an increase in the uncertainty variance with discharge, time auto-correlation, skewness, etc. Some

studies (e.g., Yang et al., 2007b; Schoups and Vrugt, 2010) account for these properties for the calibration of hydrological models within a Bayesian framework, using specific formulations of likelihood. In an extrapolation context, it is of utter importance that the predictive uncertainty assessment provides a correct description of the ~~heteroscedasticity,~~ evolution of the predictive distribution properties with the discharge magnitude. Bremnes (2019) showed that the skewness of wind speed distribution depends on the forecasted wind. Modelled residuals of discharge forecasts often exhibit high heteroscedasticy (Yang et al., 2007a). McInerney et al. (2017) focused their study on representing error heteroscedasticity of discharge forecasts with respect to simulated streamflow. To achieve reliable forecasts, a correct description of the heteroscedasticity, either explicitly or implicitly, is necessary.

Various approaches for uncertainty assessment have been developed to assess the uncertainty in hydrological predictions (see e.g., Montanari, 2011). The first step consists in identifying the different sources of uncertainty or at least the most important ones that have to be taken into account given a specific context. In the context of flood forecasting, decomposing the total uncertainty into its two main components is now common: the input uncertainty (mainly the meteorological forecast uncertainty) and the modelling uncertainty, as proposed by Krzysztofowicz (1999). More generally, the predictive uncertainty due to various sources may be explicitly modelled and propagated through the modelling chain, while the "remaining" uncertainty (from the other sources) may then be assessed by statistical post-processing.

1.2.1 Modelling each source of uncertainty

A first approach that intends to model each source of uncertainty separately and to propagate these uncertainties through the modelling chain ~~(Renard et al., 2010). The~~ is presented by Renard et al. (2010). Following 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).

In particular, the ensemble approaches intend to account for meteorological forecast uncertainty. They are increasingly popular in the research and the operational forecasting communities. An increasing number of hydrological ensemble forecasting systems are in operational use and have proved their usefulness, e.g. the European Flood Awareness System (EFAS: Ramos et al., 2007; Thielen et al., 2009; Pappenberger et al., 2011, 2016) and the Hydrologic Ensemble Forecast Service (HEFS: e.g. Demargne et al., 2014).

Multi-model approaches can be used to assess modelling uncertainty (Velazquez et al., 2010; Seiller et al., 2017). While promising, this approach requires the implementation and the maintenance of a large number of models, which can be burdensome in operational conditions. There is no evidence that such an approach ensures that the heteroscedasticity of the predictive uncertainty distribution would be correctly assessed.

In forecasting mode, data assimilation schemes based on statistical modelling are of common use to reduce and assess the predictive uncertainty. Some algorithms such as particle filters (Moradkhani et al., 2005a; Salamon and Feyen, 2009; Abbaszadeh et al., 2018) or the ensemble Kalman filter (Moradkhani et al., 2005b) provide an assessment of the predictive

uncertainty as a direct result of data assimilation ("in the loop"). Some of these approaches can explicitly account for the desired properties of the predictive uncertainty distribution, such as heteroscedasticity, through the likelihood formulation.

1.2.2 Post-processing approaches

Alternatively, numerous post-processors of deterministic or probabilistic models have been developed to account for the uncertainty from sources that are not modelled explicitly. They differ in several aspects (see a recent review by Li et al., 2017). Most approaches are conditional: the predictive uncertainty is modelled with respect to a predictor, which most often is the forecasted value (Todini, 2007, 2009). Some methods are based on predictive distribution modelling, while others can be described as "distribution-free", as mentioned by Breiman (2001). Among the former, many approaches are built in a statistical regression framework to assess the total or remaining predictive uncertainty. Examples are the Hydrologic Uncertainty Processor (HUP) in a Bayesian forecasting system (BFS) framework (Krzysztofowicz, 1999; Krzysztofowicz and Maranzano, 2004), the Model Conditional Processor (MCP: Todini, 2008; Coccia and Todini, 2011; Barbetta et al., 2017), the meta-Gaussian model of Montanari and Grossi (2008) or the Bayesian joint probability (BJP) method (Wang et al., 2009), among others. The latter approaches build a description of the predictive residuals from past error series, such as data learning-algorithms (Solomatine and Shrestha, 2009). Some related methods are the non-parametric approach of Van Steenbergen et al. (2012), the empirical hydrological uncertainty processor of Bourgin et al. (2014) or the k-nearest neighbours method of Wani et al. (2017). The Quantile Regression (QR) framework (Weerts et al., 2011; Dogulu et al., 2015; Verkade et al., 2017) lies in between in that it introduces an assumption of a linear relationship between the forecasted discharge and the quantiles of interest.

~~These approaches~~

1.2.3 Combining different approaches

The approaches presented in subsections 1.2.1 and 1.2.2 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 (Zalachori et al., 2012; Hemri et al., 2015; Abaza et al., 2017; Sharma et al., 2018). Thus, many operational flood forecasting services use post-processing techniques to assess hydrological modelling uncertainty, while meteorological uncertainty is taken into account separately (Berthet and Piotte, 2014). Post-processors are then trained with "perfect" future rainfall (i.e., equal to the observations). Moreover, even for assessing modelling uncertainty, ~~using several methodologies together may allow one to combine their respective strengths~~combining the strengths of several methodologies may improve the results.

Note that many of these approaches use a variable transformation to handle the heteroscedasticity and more generally the evolution of the predictive distribution properties with respect to the forecasted discharge. Some are non-parametric, while others use a few parameters, allowing more flexibility in the predictive distribution assessment. More details on commonly used variable transformations are presented in Section 2.1.4.

1.3 Scope

In this article, we focus on uncertainty assessment with a post-processing approach based on residuals modelling. Del Giudice et al. (2013) and McInerney et al. (2017) presented interesting comparisons of different variable transformations used for residuals modelling. Yet, their studies do not focus on the extrapolation context. Since ~~it is not possible to achieve~~ achieving a reliable predictive uncertainty assessment in an extrapolation context ~~if heteroscedasticity~~ is a challenging task likely to remain imperfect if the stability of the characteristics of the predictive distributions is not properly ~~taken into account, the~~ ensured, it requires a specific crash-testing framework (Andréassian et al., 2009). The objectives of this article are:

- to present a framework aimed at testing the hydrological modelling and uncertainty assessment in the extrapolation context;
- to assess the ability and the robustness of a post-processor to provide reliable predictive uncertainty assessment for large floods when different variable transformations are used;
- to provide guidance for operational flood forecasting system development.

We attempt to answer ~~two~~ three questions: (a) Can we improve residuals modelling with an adequate variable transformation in an extrapolation context? (b) Do more flexible transformations, such as the log-sinh transformation, help in obtaining more reliable predictive uncertainty assessment? (c) If the performance decreases when extrapolating, is there any driver that can help the operational forecasters to predict this performance loss and question the quality of the forecasts ?

Section 2 describes the data, the forecast model, the post-processor and the testing methodology chosen to address these questions. Section 3 presents the results of the numerical experiments that are then discussed in ~~Sections 4 and 5.~~ Section 4. Finally, a number of conclusions and perspectives are proposed.

2 Data and methods

2.1 Data and forecasting model

2.1.1 Catchments and hydrological data

We used a set of 154 unregulated catchments spread throughout France (Fig. 1) ~~to test our hypotheses~~ over various hydrological regimes and forecasting contexts to provide robust answers to our research questions (Andréassian et al., 2006; Gupta et al., 2014). They represent a large variability in climate, topography and geology in France (Table 1), although their hydrological regimes are little or not at all influenced by snow accumulation. Hourly rainfall, potential evapotranspiration (PE) and streamflow data series were available over the 1997 – 2006 period. PE was estimated using a temperature-based formula (Oudin et al., 2005). Rainfall and temperature data come from a radar-based reanalysis produced by Météo-France (Tabary et al., 2012). Discharge data were extracted from the national streamflow HYDRO archive (www.hydro.eaufrance.fr). When a hydrological model is used to issue forecasts, it is often necessary to compare the lead-time to a characteristic time of the catchment (section 2.1.2). For each catchment, the lag time LT is estimated as the lag time maximising the cross-correlation between rainfall and discharge time series.

Table 1. Characteristics of the 154 catchments, computed over the 1997 – 2006 data series.

	~~Percentiles~~Quantiles						
	0	0.05	0.25	0.50	0.75	0.95	1
Catchment area (km^2)	9	27	79	184	399	942	3,260
Average altitude (m above sea level)	64	92	188	376	589	897	1,050
Average slope (%)	2	3	6	9	18	32	39
Lag time (h)	3	5	9	12	19	29	33
Mean annual rainfall (mm/yr)	639	727	876	1,003	1,230	1,501	1,841
Mean annual potential evapotranspiration (mm/yr)	549	549	631	659	700	772	722
~~Mean~~ Specific mean annual discharge (mm/yr)	53	142	262	394	583	1,114	1,663
Mean annual discharge (m^3/s)	1	1	1	2	5	17	53
Maximum hourly rainfall (mm/h)	10	12	16	20	26	41	61
Quantile 0.99 of the hourly discharge (m^3/s)	1	2	6	15	33	115	296

2.1.2 Hydrological model

We used discharge forecasts computed by the GRP rainfall-runoff model. The GRP model is designed for flood forecasting and is currently used by the Flood Forecasting Services (FFS) in France in operational conditions (Furusho et al., 2016; Viatgé et al., 2018). It is a deterministic lumped storage-type model that uses catchment areal rainfall and PE as inputs. ~~The~~ In

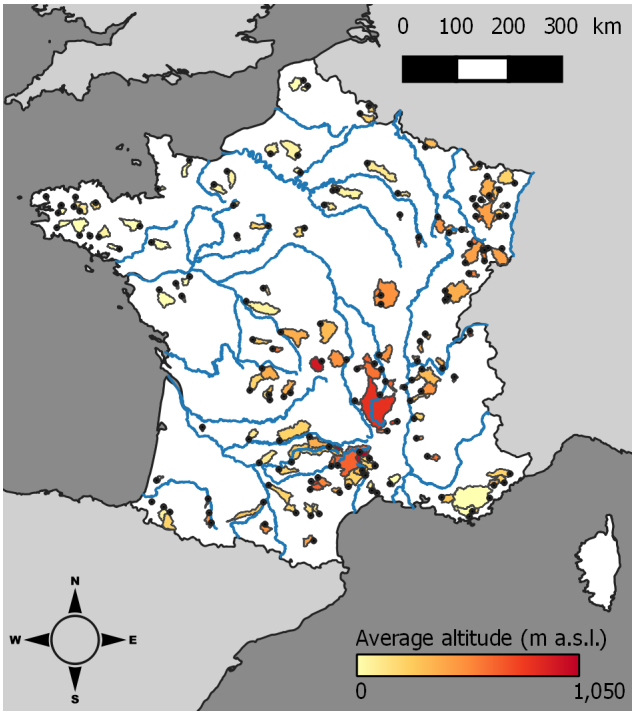

Figure 1. The set of 154 unregulated catchments used in this study. Average altitude is given in meters above sea level (m a.s.l.).

forecasting mode (Appendix A), the model also assimilates discharge observations available when issuing a forecast to update the main state variable of the routing function and to update the output discharge. In this study, it is run at the hourly time step and forecasts are issued for several lead times ranging from 1 to 72 h. More details about the GRP model can be found in Appendix A.

Since herein only the ability of the post-processor to extrapolate uncertainty quantification is studied, the model is fed only with observed rainfall (no forecast of precipitation), in order to reduce the impact of the input uncertainty. For the same reason, 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. The results will be presented for four lead times, LT / 2, LT , 2 LT and 3 LT, to cover the different behaviours that can be seen when data assimilation is used to reduce errors in an operational flood forecasting context (Berthet et al., 2009) .

2.1.3 Empirical hydrological uncertainty processor (EHUP)

We used the empirical hydrological uncertainty processor (EHUP) presented in Bourgin et al. (2014). It is a data-based and non-parametric approach to estimate the conditional predictive uncertainty distribution. This post-processor was compared to other post-processors in earlier studies and proved to provide relevant results (Bourgin, 2014). It is now used by operational FFS in France under the operational tool called OTAMIN (Viatgé et al., 2019). ~~Separately for each lead-time,~~ The main difference

with many other post-processors (such as the MCP, the meta-Gaussian processor, the BJP or the BFS) is that no assumption is made about the shape of the uncertainty distribution, which brings more flexibility to represent the various forecast error characteristics encountered in large sample modelling. We will discuss the impact of this choice in Sect. 4.2.

The basic idea of the empirical hydrological uncertainty processor (EHUP) is to estimate empirical quantiles of errors ~~are estimated for~~ stratified by different flow groups ~~obtained by~~ to account for the variation of the forecast error characteristics with forecast magnitude. Since forecast error characteristics also vary with the lead time when data assimilation is used, the EHUP is trained separately for each lead time.

For each lead time separately, the following steps are used:

1. Training:
   - The flow groups are obtained by first ordering the forecast-observation pairs according to the forecasted values ~~.~~ and then stratifying the pairs into a chosen number of groups (in this study, we used 20 groups), so that each group contains the same number of pairs.
   - Within each flow group, errors are calculated as the difference between the two values of each forecast-observation pairs and several empirical quantiles (we used 99 percentiles) are calculated in order to characterize the distribution of the error values.
2. Application:
   - The predictive uncertainty distribution that is associated to a given (deterministic) forecasted value is defined by adding this forecasted value to the empirical quantiles that belong to the same flow group as the forecasted value.

Since we focus ~~here~~ on the extrapolation case, only the highest flow group ~~containing~~ is used in this study. Here, this highest flow group contains the top 5% pairs ranked by ~~simulated valuesis used~~forecasted values. This threshold is chosen as a compromise between focusing at the highest values and using a sufficiently large number of forecast-observation pairs when estimating empirical quantiles of errors. ~~Heteroscedasticity of the extrapolated predictive distributions is taken into account using different data transformations described in the next subsection~~In extrapolation, when the forecast discharge is higher than the highest value of the training period, the predictive distribution of the error is kept constant, i.e., the same values of the empirical quantiles of errors are used, as illustrated in Fig. 5.

The EHUP can be applied after a preliminary data transformation, and by adding a final step to back-transform the predictive distributions obtained in a transformed space. In previous work, we used the log transformation because it ensures that no negative values are obtained when estimating the predictive uncertainty for low flows. When estimating the predictive uncertainty for high flows, the data transformation has a strong impact in extrapolation, because the variation of the extrapolated predictive distribution, which is constant in the transformed space, is controlled in the real space by the behaviour of the inverse transformation, as explained below.

2.1.4 The different transformation families

Many uncertainty assessment methods mentioned in the introduction use a variable transformation to handle the heteroscedasticity of the residuals and account for the variation of the prediction distributions with the magnitude of the predicted variable. Here, we briefly recall a number of variable transformations commonly used in hydrological modelling. Let y and $\tilde{y}$ be the observed and forecasted variables (here, the discharge) and $\varepsilon = y - \tilde{y}$ the residuals. When using a transformation g, we consider the residuals $\varepsilon' = g(y) - g(\tilde{y})$.

~~The normal quantile transformation (NQT) is a non-parametric transformation linking a given distribution and the Gaussian distribution, quantile by quantile:~~

$$g_1(y) = \mathrm{NQT}(y) = \phi^{-1}(F(y))$$

~~where ϕ is the normal Gaussian cumulative density function (cdf) and F is the cdf of a variable Y. When used in a post-processor, F is most often empirically computed from a large sample of residuals. While several hydrological processors such as the HUP, MCP and the Quantile Regression encompass the NQT-transformed variables, Bogner et al. (2012) warn against the drawbacks of this transformation, which is by construction not suited for the extrapolation context and requires additional assumptions to model the tails of the distribution (Weerts et al., 2011; Coccia and Todini, 2011).~~

~~The next three transformationsare analytical~~Three analytical transformations are often met in hydrological studies: the log, Box-Cox and log-sinh transformations. The log transformation is commonly used (e.g., Morawietz et al., 2011):

$$g_2(_1 : y) = \mapsto \log(y + a)$$

where a is a small positive constant to deal with y values close to 0. It can be taken equal to 0 when focusing on large discharge values. This transformation is then non-parametric (no parameter has to be calibrated). Applying a statistical model on residuals computed on log-transformed variables may be interpreted as using a corresponding model on multiplicative error (e.g., assuming a Gaussian model for residuals of log-transformed discharges is equivalent to a log-Normal model on the multiplicative errors $y/\tilde{y}$). Therefore, it may be adapted to strongly heretoscedastic behaviours. It has been used successfully to assess hydrological uncertainty (Yang et al., 2007b; Schoups and Vrugt, 2010; Bourgin et al., 2014).

The Box-Cox transformation (Box and Cox, 1964) is a classic one-parameter transformation that is quite popular in the hydrological community (e.g., Yang et al., 2007b; Wang et al., 2009; Hemri et al., 2015; Reichert and Mieleitner, 2009; Singh et al., 2013; Del Giudice et al., 2013):

$$g_3(_{2\ [\lambda]} : y) = \mapsto \frac{(y+a)^\lambda - 1}{\lambda} \quad \text{if} \quad \lambda \neq 0 \quad \text{and} \quad g_3(y)_{2\ [\lambda]} = g_2(y)_1 \quad \text{if} \quad \lambda = 0$$

The Box-Cox transformation makes it possible to cover very different behaviours. The log transformation is a special case of the Box-Cox transformation when the calibration results in $\lambda = 0$. In contrast, applying the Box-Cox transformation with $\lambda = 1$ on the variable y to model the distribution of their residuals is equivalent to applying no transformation (Fig. 2). McInerney et al. (2017) obtained their ~~best~~ most reliable and sharpest results with $\lambda = 0.2$ over 17 perennial catchments.

More recently, the log-sinh transformation has been proposed (Wang et al., 2012; Pagano et al., 2013). It is a two-parameter transformation:

$$g_{\alpha,\beta 3\ [\alpha,\beta]} : y \mapsto \beta \cdot \log\left(\sinh\left(\frac{\alpha + y}{\beta}\right)\right)$$

This transformation provides more flexibility. Indeed, for $y \gg \alpha$ and $y \gg \beta$, the log-sinh transformation reduces to no transformation, while for $\alpha \ll y \ll \beta$, it is equivalent to the log transformation (Fig. 2). Thus, with the same ~~parametrization~~parameterization, it can result in very different behaviours depending on the magnitude of the discharge. Applying no transformation may be intuitively attractive to model the residuals distribution for ~~very~~ large discharge values, when the variance is not longer expected to increase (homoscedastic behaviour). It is then particularly attractive when modelling predictive uncertainty in an extrapolation context, in order to avoid an excessively "explosive" assessment of the predictive uncertainty for large discharge values.

In addition to the log transformation used by Bourgin et al. (2014), in this study we tested the Box-Cox and the log-sinh transformations to explore more flexible ways to deal with the challenge of extrapolating prediction uncertainty distributions (Fig. 2). The impacts of the data transformations used in this study are illustrated in Fig. 3.

Another common variable transformation is the normal quantile transformation (NQT). It is a non-parametric transformation linking a given distribution and the Gaussian distribution, quantile by quantile:

$$g_4 : y \mapsto \mathrm{NQT}(y) = \phi^{-1}(F(y))$$

where ϕ is the normal Gaussian cumulative density function (cdf) and F is the cdf of a variable Y. When used in a post-processor, F is most often empirically computed from a large sample of residuals. While several hydrological processors such as the HUP, MCP and the Quantile Regression encompass the NQT-transformed variables, Bogner et al. (2012) warn against the drawbacks of this transformation, which is by construction not suited for the extrapolation context and requires additional assumptions to model the tails of the distribution (Weerts et al., 2011; Coccia and Todini, 2011). This is why we did not test this transformation in this study focused on the extrapolation context.

2.2 Methodology: a testing framework designed for extrapolation context assessment

2.2.1 Testing framework

The ~~EHUP post-processor~~ empirical hydrological uncertainty processor (EHUP) is a non-parametric approach based on the characteristics of the residuals distribution over a training data set. Moreover, the Box-Cox and the log-sinh transformations are parametric and require a calibration step. Therefore, the methodology adopted for this study is a split-sample scheme test inspired by the differential split-sample scheme of Klemeš (1986) and based on three data subsets: a data set for training the EHUP, a data set for calibrating the parameters of the variable transformation, and a control data set for evaluating the predictive distributions when extrapolating high flows.

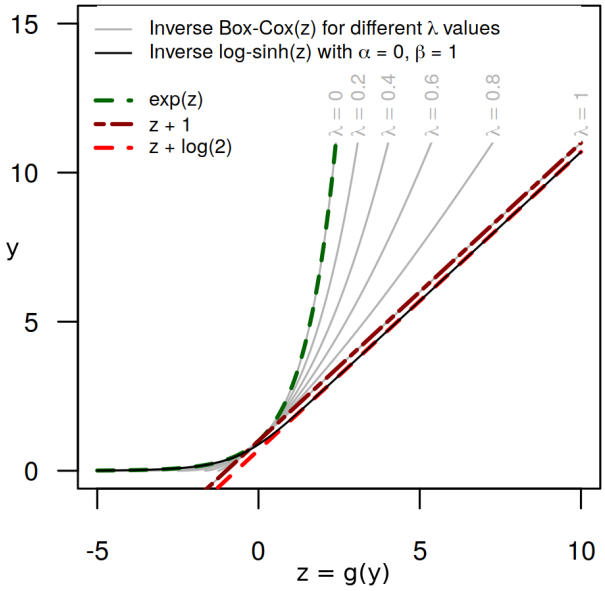

Figure 2. ~~Left panel~~The inverse transformation explains the final effect on the uncertainty assessment: the ~~Box-Cox transformation for various parameter values (grey straight lines) and~~ constant probability distribution in the ~~log-sinh transformation for one parametrization~~ transformed space (~~black straight line~~provided by the EHUP) ~~. Right panel:~~ will result in a distribution in the untransformed space, whose evolution depends on the behaviour of the ~~corresponding~~ inverse ~~transformations~~data transformation. ~~The~~ Here, the Box-Cox transformation provides different behaviours, depending on its parameter value (λ). It ranges from an affine transformation (equivalent to no transformation; red dashed line) to the log transformation (thick green straight line). With a single ~~parametrization~~parameterization, the log-sinh transformation can be equivalent to the log transformation for values of y much smaller than the value of its parameter β and equivalent to an affine transformation for large values of y (much higher than β; see Appendix B).

2.2.2 Events selection

To populate the three data subsets with independent data, separate flood events were first selected by an iterative procedure similar to those ~~presented in~~detailed by Lobligeois et al. (2014) and Ficchì et al. (2016): (1) the maximum forecasted discharge of the time series was selected, (2) within a 20-day period before (after) the peak flow, the beginning (end) of the event was placed at the ~~first time step~~ preceding (following) time step closest to the peak at which the streamflow is lower than 20% (25%) of the peak flow value, (3) the event was kept if there was less than 10% missing values, if the beginning and end of the event were lower than 66% of the peak flow value and if the peak value was higher than 50% of the highest discharge value of the time series. The process is then iterated over the remaining data to select all events. A minimum time lapse of 24 h was enforced between two events, ensuring that consecutive events are not overlapping and that the autocorrelation between the time steps of two separate events remains limited.

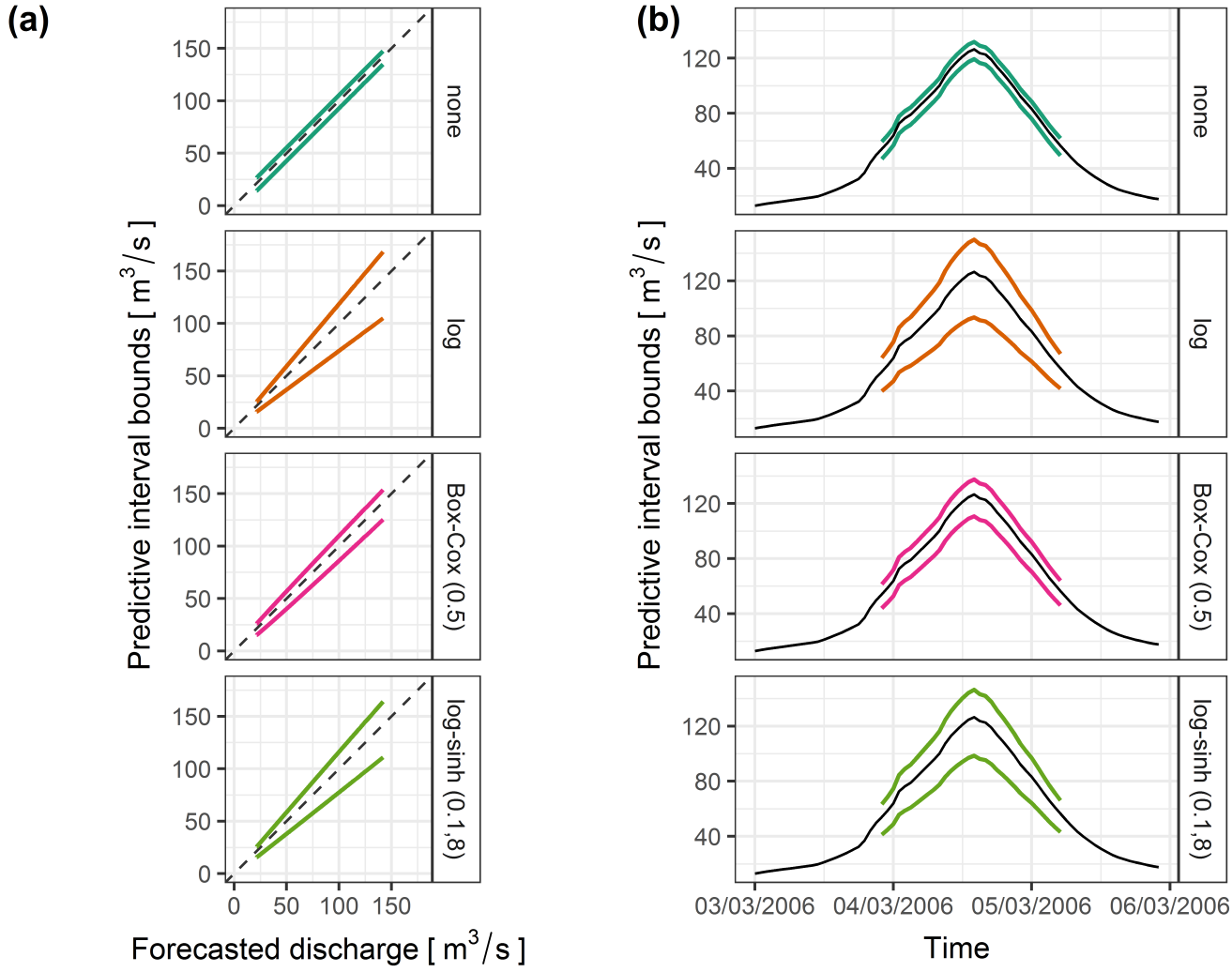

Figure 3. Predictive 0.1 and 0.9 quantiles when assessed with no transformation, the log-transformation, the Box-Cox transformation with its λ parameter equal to 0.5 and the log-sinh with its α and β parameters equal to 0.1 and 8, on the Ill River at Didenheim (668 km^2) for lead time LT: (a) as a function of the deterministic forecasted discharge and (b) for the flood event on 3 March 2006. The heteroscedasticity strongly differs from one variable transformation to another.

The number of events and their characteristics vary greatly among catchments, as summarised in Table 2. Note that the events selected for one catchment can slightly differ for the four different lead times considered in this study, because the selection was made using the forecasted discharge and not the observed discharge.

Table 2. Characteristics of the events selected for the lead time LT over the 1997 – 2006 data series.

	~~Percentiles~~Quantiles						
	0	0.05	0.25	0.50	0.75	0.95	1
Total length of events (days)	663	1178	1299	1505	1808	2061	27
Number of events G1	28	65	114	141	193	276	43
Number of events G2	8	16	23	32	43	58	17
Number of events G3	7	12	19	26	35	46	16
~~Median~~ Specific median value of the peak discharges G1 (~~m~~^{3}mm/~~s~~h)	~~1~~ 0.009	~~1~~ 0.022	~~1~~ 0.041	~~3~~ 0.062	~~7~~ 0.098	~~11~~ 0.184	~~87~~ 0.
~~Median~~ Specific median value of the peak discharges G2 (~~m~~^{3}mm/~~s~~h)	~~1~~ 0.026	~~2~~ 0.064	~~5~~ 0.140	~~11~~ 0.217	~~24~~ 0.342	~~49~~ 0.706	~~254~~ 1
~~Median~~ Specific median value of the peak discharges G3 (~~m~~^{3}mm/~~s~~h)	~~2~~ 0.050	~~3~~ 0.131	~~9~~ 0.295	~~23~~ 0.439	~~53~~ 0.701	~~103~~ 1.896	~~502~~ 3

2.2.3 Selection of the data subsets

The selected events were then gathered into three events sets, G1, G2 and G3, based on the magnitude of their peaks and the number of useful time steps for each test phase (training of the EHUP post-processor, calibration of the variable transformations, evaluation of the predictive distributions): G1 contains the lowest events, while the highest events are in G3.

The selection of the data subsets was tailored to study the behaviour of the post-processing approach in an extrapolation context. The control data subset had to encompass only time steps with simulated discharge values higher than those met during the training and calibration steps. Similarly, the calibration data subset had to encompass time steps with simulated discharge values higher than those of the training subset.

To achieve these goals, only the time steps within flood events were used. We distinguished four data subsets, as illustrated in Fig. 4. The subset D1 gathered all the time steps of the events of the G1 group. Then, the set D2 of the time steps of the events of the G2 group was split into two subsets: $D2_{sup}$ gathered all the time steps with forecasted discharge values higher than the maximum met on D1, and $D2_{inf}$ was filled with the other time steps. Finally, D3 was similarly filled with all the time steps of the G3 events with forecasted discharge values higher than the maximum met on D2.

The discharge thresholds used to populate the D1, $D2_{sup}$ and D3 subsets from the events belonging to the G1, G2 and G3 groups were chosen to ensure a sufficient number of time steps in every subset. We chose to set the minimum number of time steps in D3 and $D2_{sup}$ to 720 as a compromise between having enough data to evaluate the methods and keeping the extrapolation range sufficiently large. We lowered this limit to 500 for the top 5% pairs of D1, since this subset was only used to build the empirical distribution by estimating 99 percentiles during the training step and not used for evaluating the quality of the uncertainty assessment.

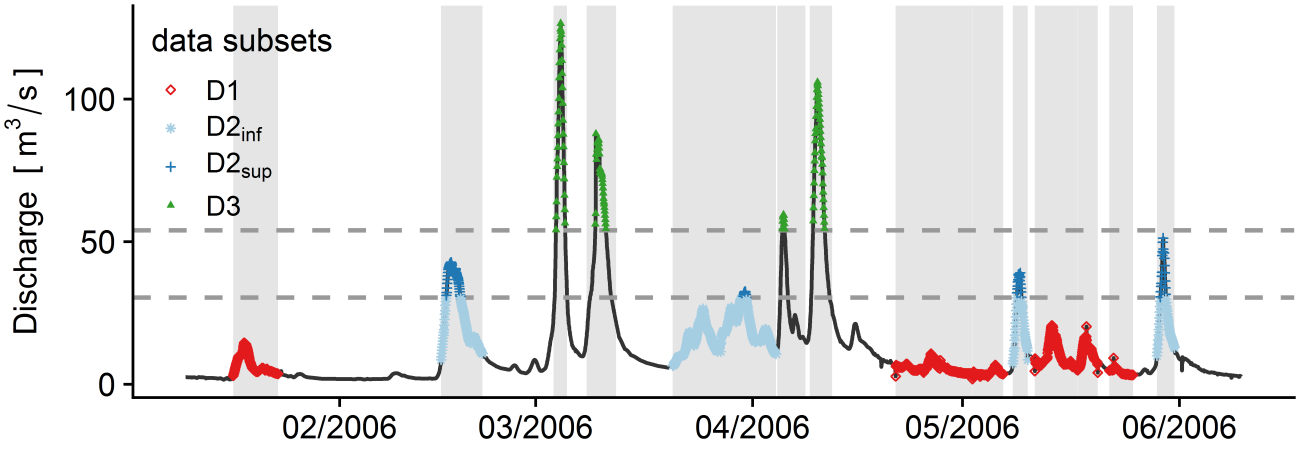

Figure 4. Illustration of ~~(a)~~ the selection of ~~events in grey and (b)~~ the ~~selection of the four~~ data subsets for the Ill River at Didenheim (668 km^2[2]). ~~The horizontal dashed lines show~~ First, the ~~thresholds between~~ events are selected (grey highlighting). Then, the four data subsets are populated according to the thresholds (horizontal dashed lines). See Sect. 2.1.3 for more details.

2.2.4 Calibration and evaluation steps

Since there are only one parameter for the Box-Cox transformation and two parameters for the log-sinh transformation, a simple calibration approach of the transformation parameters was chosen: the parameter space was explored by testing several parameter set values. For the Box-Cox transformation, 17 values for the λ parameter were tested: from 0 to 1 with a step of 0.1 and with a refined mesh for the sub-intervals 0 – 0.1 and 0.9 – 1. For the log-sinh transformation, a grid of 200 (α, β) values was designed for each catchment based on the maximum value of the forecasted discharge on the D2 subset, as explained in greater detail in the Appendix B.

Note that the hydrological model was calibrated over the whole set of data (1997 – 2006) to make the best use of the data set, since this study focuses on the effect of extrapolation on the predictive uncertainty assessment only.

We used a two-step procedure, as illustrated in Fig. 5. In the first step, for each transformation and for each parameter set, the empirical residuals were computed over D1 and the calibration criterion was computed on $D2_{sup}$. Indeed, the data transformations have almost no impact on the uncertainty estimation by EHUP on events of the same magnitude as those of the training subset. Therefore the calibration subset has to encompass events of a larger magnitude ($D2_{sup}$). The parameter set obtaining the best criterion value was selected. In the second step, the ~~empirical residuals distribution was estimated~~ EHUP was trained on a data set which encompassed D1, $D2_{inf}$ and $D2_{sup}$ using the parameter set obtained during the calibration step~~, and~~ . Then, the predictive uncertainty distribution was evaluated on the control data set D3. ~~This second estimation of the residuals distribution made it possible to test the model~~ Training the EHUP on the union of D1, $D2_{inf}$ and $D2_{sup}$ allows to control the

uncertainty assessment from small to large degrees of extrapolation ~~, on~~ (on D3). Indeed if we had kept the training on D1 only, we would have not been able to test small degrees of extrapolation on independent data for every catchment (see the discussion in Sect. 3.3).

2.3 Performance criteria and calibration

2.3.1 Probabilistic evaluation framework

Reliability was first assessed by a visual inspection of the Probability integral transform (PIT) diagrams (Laio and Tamea, 2007; Renard et al., 2010). Since this study was carried out over a large sample of catchments, two standard numerical criteria were used to summarise the results: the α-index, which is directly related to the PIT diagram (Renard et al., 2010), and the coverage rate of the 80% predictive intervals (bounded by the 0.1 and 0.9 quantiles of the predictive distributions), used by the French operational FFS. The α-index is equal to $1 - 2 \cdot A$, where A is the area between the PIT curve and the bisector, and its value ranges from 0 to 1 (perfect reliability).

The overall quality of the probabilistic forecasts was evaluated with the Continuous Rank Probability Score (CRPS, Hersbach, 2000), which compares the predictive distribution to the observation:

$$\mathrm{CRPS} = \frac{1}{N} \sum_{k=1}^{N} \int_{0}^{\infty + \infty} \left[F_k(Q) - H(Q - Q_{k,obs}) \right]^2 dQ$$

where N is the number of time steps, $F(Q)$ is the predictive cumulative distribution, H the Heaviside function and Q_{obs} is the observed value. We used a skill score (CRPSS) to compare the mean CRPS to a reference, here the mean CRPS obtained from the unconditional climatology, i.e. from the distribution of the observed discharges over the ~~events selected~~same data subset.

For operational purposes, the sharpness of the probabilistic forecasts was checked by measuring the mean width of the 80% predictive intervals. A ~~non-dimensional~~ dimensionless relative sharpness index was obtained by dividing the mean width by the mean runoff:

$$1 - \frac{\sum_{k=1}^{N} \left[q_{0.9}(Q_k) - q_{0.1}(Q_k) \right]}{\sum_{k=1}^{N} Q_{k,obs}}$$

where $q_{90}(Q)$ and $q_{10}(Q)$ are the upper and the lower ~~bound~~ bounds of the 80% predictive interval for each forecast.

~~Finally~~In addition to the probabilistic criteria presented above, the accuracy of the forecasts was assessed using the Nash-Sutcliffe Efficiency (NSE) calculated with the mean values of the predictive distributions.

2.3.2 The calibration criterion

Since the calibration step aims at selecting the most reliable description of the residuals in extrapolation, the α-index was used to select the parameter set that yields the highest reliability for each catchment, each lead time and each transformation. While other choices were possible, we followed the paradigm presented by Gneiting et al. (2007): reliability has to be ensured before

sharpness. Note that the CRPS could have been chosen, since it can be decomposed as the sum of two terms: reliability and sharpness (Hersbach, 2000). However, in the authors' experience, the latter is the controlling factor (Bourgin et al., 2014). Moreover, the CRPS values were often quite insensitive to the values of the log-sinh transformation parameters.

In cases where an equal value of the α-index was obtained, we selected the parameter set that gave the best sharpness index. For the log-sinh transformation, there were still a few cases where an equal value of the sharpness index was obtained, revealing the lack of sensitivity of the transformation in some areas of the parameter space. For those cases, we chose to keep the parameter set that had the lowest ~~γ_1~~α value and the ~~γ_2~~β value closest to ~~1.~~ $max_{D2}\left(\widetilde{Q}\right)$.

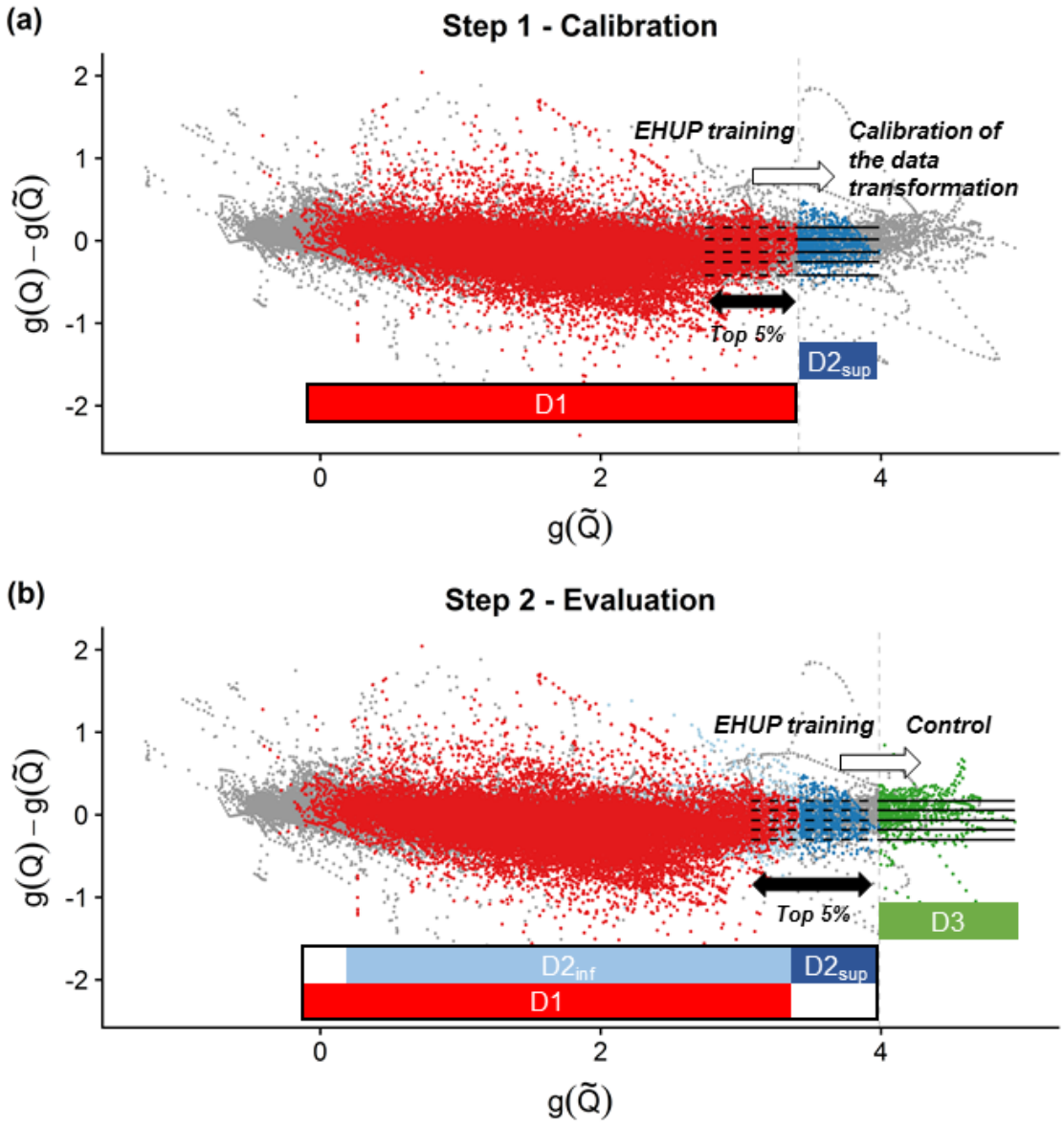

Figure 5. Residuals as a function of the forecast discharges in the transformed space. The horizontal dashed lines represent the 0.1, 0.25, 0.5, 0.75 and 0.9 quantiles of the residuals computed during the training phase of the EHUP post-processor, ~~while~~ for the highest flow group (top 5% pairs of the training data ranked by forecasted values). The straight lines represent their use to assess the predictive uncertainty in extrapolation during (a) the calibration step of the variable transformation parameters and (b) the evaluation step of the predictive uncertainty. The vertical dashed lines show the beginning of the extrapolation range. Illustration from the Ill River at Didenheim, 668 km^2. Data used for the EHUP training at each step is sketched by a thick rectangle. In this study, only the top 5% pairs of the training data ranked by forecasted values is used to estimate the residual distribution, since we focus on the extrapolation behaviour of the EHUP.

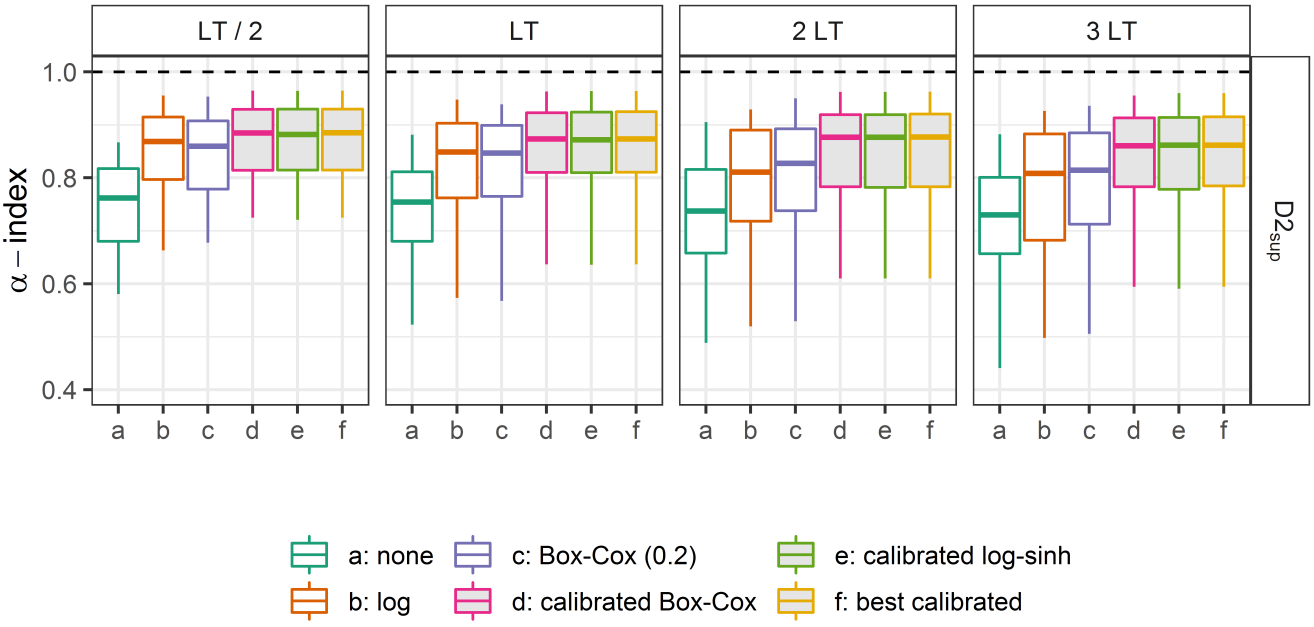

Figure 6. Distributions of the α-index values on the calibration data set $D2_{sup}$, obtained with different transformations for four lead times (the filled boxplots represent the calibrated distributions). Boxplots (5th, 25th, 50th, 75th and 95th percentiles) synthesise the variety of scores over the catchments of the data set.

3 Results

3.1 Results on the calibration data set $D2_{sup}$

Figure 6 shows the distributions of the α-index values obtained with different transformations on the calibration data set ($D2_{sup}$) for lead times LT / 2, LT, 2 LT, 3 LT. The distributions are summarised with boxplots. Clearly, not using any transformation leads to poorer reliability than any tested transformation. In addition, we note that the calibrated transformations provide better results (although not perfect) than the non-parametric ones on the calibration data set, as expected, and that ~~there is no significant difference~~ no noticeable difference can be seen in Fig. 6 between the calibrated Box-Cox transformation (d), the calibrated log-sinh transformation (e) and the best ~~performing~~ calibrated transformation (f). Nevertheless, the uncalibrated log transformation and Box-Cox transformation with parameter λ set at 0.2 ($BC_{\lambda=0.2}$) reach quite reliable forecasts. ~~Interestingly, the log transformation provides the best results for the other criteria (not used as the objective function).~~ Comparing the results obtained for the different lead times reveals that less reliable predictive distributions are obtained for longer lead times, in particular for the non-parametric transformations.

Figures 7 and 8 show the distribution of parameter values obtained for the Box-Cox and the log-sinh transformation during the calibration step. The distributions vary with lead time. While the log-transformation behaviour is frequently chosen for

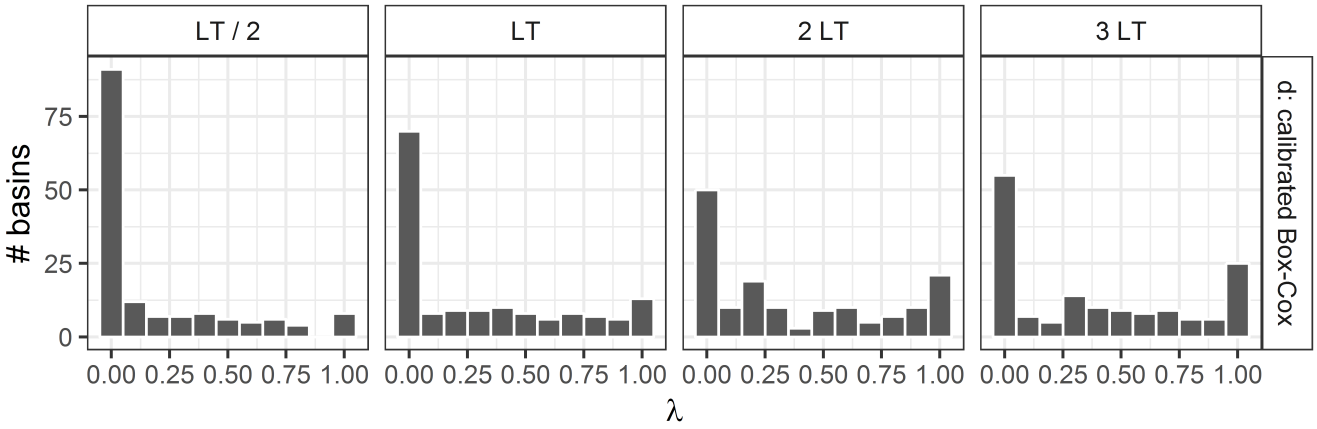

Figure 7. Distribution over the basins of the values of the Box-Cox transformation parameter obtained during the calibration step for the four different lead times.

LT/2 and LT, the additive behaviour (corresponding to the use of no transformation, see Sect. 2.1.4) becomes more frequent for 2 LT and 3 LT. A similar conclusion can be drawn for the log-sinh transformation: a low value of ~~γ_1~~ α and a high value of ~~γ_2~~ β yield a multiplicative behaviour that is frequently chosen, for all lead times, but less for 2 LT and 3 LT than for LT / 2 and LT. This explains in particular the loss of reliability that can be seen for the log transformation for LT 3 in Fig. 6. These results reveal that the extrapolation behaviour of the residuals distributions is complex. It varies among catchments and with lead time because of the strong impact of data assimilation.

3.2 Results on the D3 control data set

3.2.1 Reliability

First, we conducted a visual inspection of the PIT diagrams, which convey an evaluation of the overall reliability of the probabilistic forecasts. They show quite different patterns on the set of catchments, highlighting bias or under-dispersion problems for some of them, as illustrated in Fig. 9.

Then the distribution of the α-index values in Fig. 10 reveals a significant loss of reliability compared to the values obtained on the calibration data set (Fig. 6). We note that the log transformation is the most reliable approach for LT / 2 and is comparable to the Box-Cox transformation ($BC_{\lambda=0.2}$) for LT. With increasing lead time, the $BC_{\lambda=0.2}$ transformation becomes slightly better than the other transformations, including the calibrated ones. In addition, comparing the results obtained for the different lead times confirms that it is challenging to produce reliable predictive distributions when extrapolating at longer lead times. Overall, it means that the added value of the flexibility brought by the calibrated transformations is not transferable in an independent extrapolation situation, as illustrated in Fig. 11.

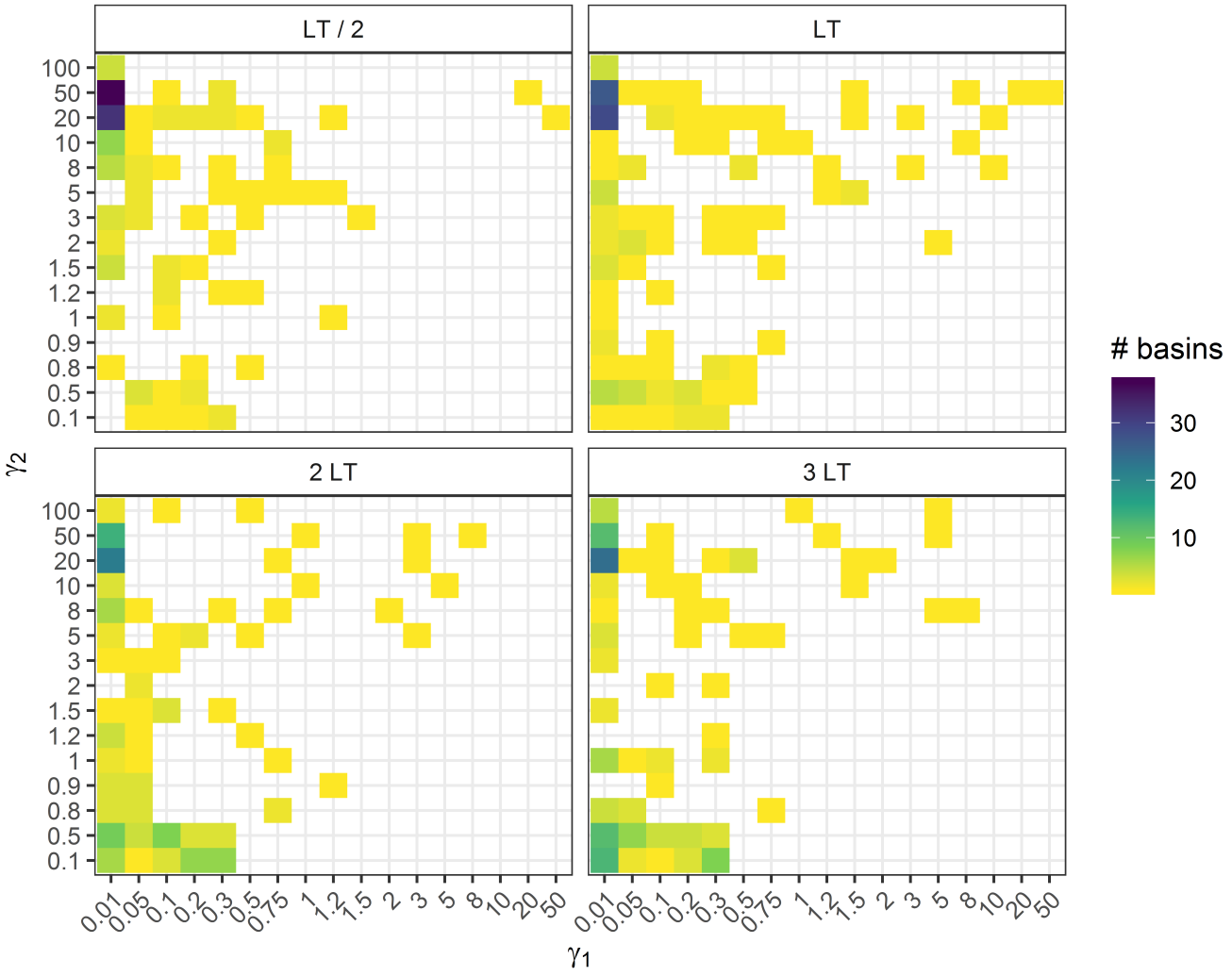

Figure 8. Distribution over the basins of the values of the log-sinh transformation parameters obtained during the calibration step for the four different lead times. $\gamma_1 = \alpha / max_{D2}\left(\tilde{Q}\right)$ and $\gamma_2 = \beta / max_{D2}\left(\tilde{Q}\right)$ (see Appendix B).

In operational settings, non-exceedance frequencies of the ~~lower (0.1 quantile) and upper (0.9 quantile)~~ quantiles of the predictive distribution which are the lower and upper bounds of the predictive interval ~~of the theoretical 80% coverage rate~~ communicated to the authorities are of particular interest~~: it~~. The 80%-predictive interval (bounded by the 0.1 and 0.9 quantiles) is mostly used in France. It is expected that ~~those values~~ the non-exceedance frequency of the lower bound and the exceedance frequency of the upper bound remain close to 10% ~~. Figures ?? and ?? reveal~~ for a reliable predictive distribution. Deviations from these frequencies indicates biases in the estimated quantiles. Figure 12 reveals that on average the 0.1 quantile is generally better assessed than the 0.9 quantile on average, ~~which is not the most desired behaviour for operational matters~~though

the latter is generally more sought for operational purposes. More importantly, ~~it can be seen that~~ the lack of reliability of the log transformation ~~seen for 3-LT~~ for the 3-LT lead time seen in Fig. 10 appears to be related to an underestimation of the ~~exceedance frequency for the~~ 0.1 quantile, ~~while the non-exceedance frequency for the~~ which is higher than for the other tested transformations, while the 0.9 quantile ~~remains limited compared to~~ is less underestimated than for the other transformations. These results highlight that reliability can have different complementary facets and that some parts of the predictive distributions can be more or less reliable. In a context of flood forecasting, particular attention should be given to the upper part of the predictive distribution.

3.2.2 Overall performance

In addition to reliability, we looked at other qualities of the probabilistic forecasts, namely the overall performance (measured by the CRPSS) and accuracy. We also checked their sharpness. The distributions of four performance criteria are showed for lead time LT in Fig. 13. We note that the log transformation has the closest median value for the coverage ratio, at the expense of a lower median relative sharpness value, because of larger predictive interval widths caused by the multiplicative behaviour of the log transformation. In addition, the CRPSS and the NSE distributions have limited sensitivity to the variable transformation (also shown by Woldemeskel et al. (2018) for the CRPS), even if we can see that not using any transformation yields slightly better results. This confirms that the CRPSS itself is not sufficient to evaluate the adequacy of uncertainty estimation. Similar results were obtained for the other lead times (~~see the figures in Supplementary Material~~Supplementary materials).

3.3 Investigating the performance loss in an extrapolation context

For operational forecasters, it is important to be able to predict when they can trust the forecasts issued by their models and when their quality becomes questionable. Therefore we investigated whether the reliability and reliability loss observed in an extrapolation context were correlated with some properties of the forecasts. First, Fig. 14 shows the relationship between the α-index values obtained in $D2_{sup}$ and those obtained in D3 for three representative transformations. The results indicate that it is not possible to anticipate the α-index values when extrapolating high flows in D3 based on the α-index values obtained when extrapolating high flows in $D2_{sup}$.

In addition, two indices were chosen to describe the degree of extrapolation: the ratio of the median of the forecasted discharges on D3 over the median of the forecasted discharge on $D2_{sup}$, and the ratio of the median of the forecasted discharges on D3 over the discharge for a return period of 20 years (for catchments where the assessment of the vicennial discharge was available in the national database: http://www.hydro.eaufrance.fr). In both cases, no trend appears, regardless of the variable transformation used, with Spearman coefficients values (much) lower than 0.33. The reliability can remain high for some catchments even when the magnitude of the events of the control data set are much higher than those of the training data set (see Supplementary materials for figures).

Finally, we found no correlation with the relative accuracy of the deterministic forecasts either. The goodness-of-fit during the calibration phase cannot be used as an indicator of the robustness of the uncertainty estimation in an extrapolation context (see Supplementary materials for figures).

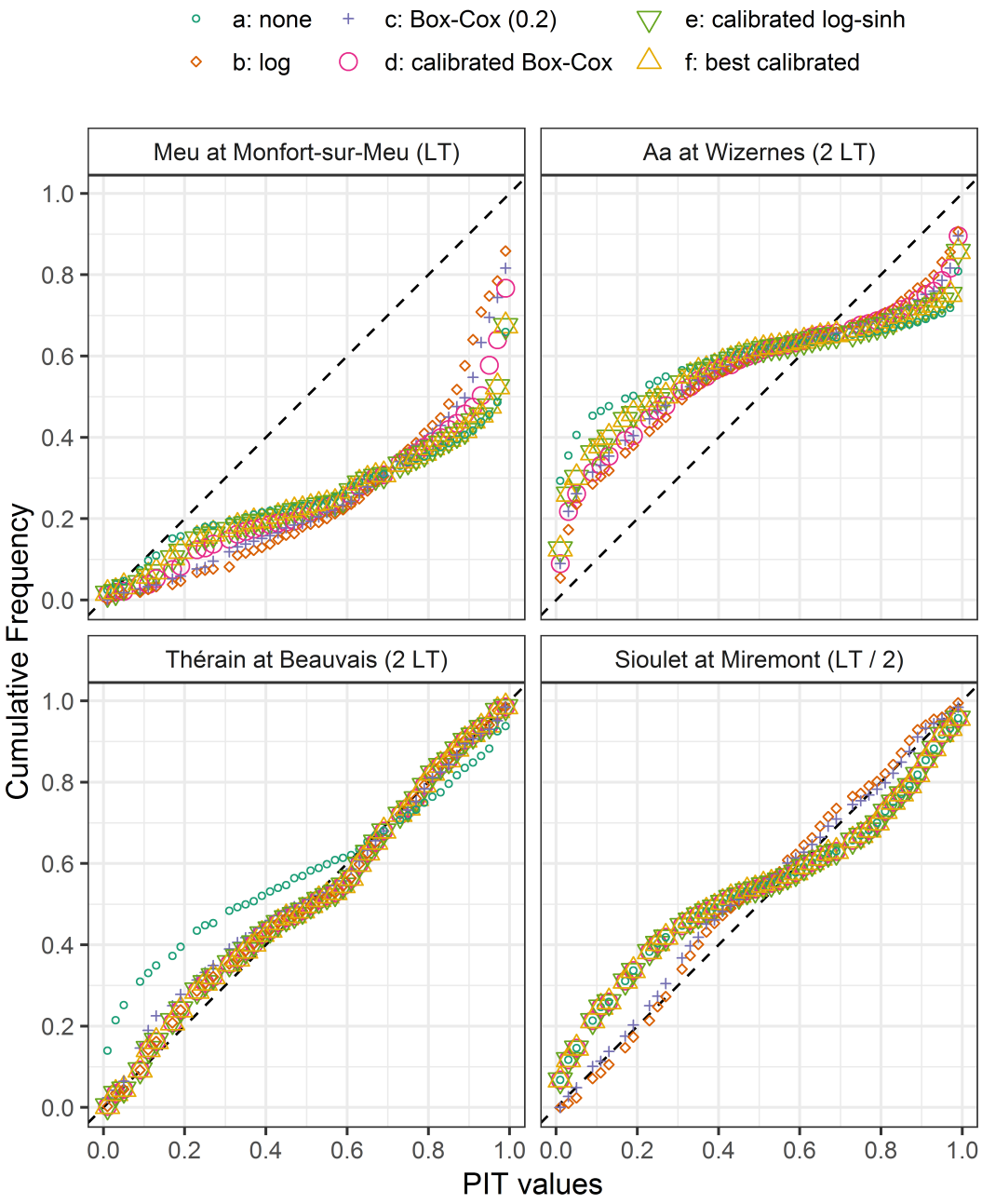

Figure 9. Examples of PIT diagrams obtained on the control data set D3, with different transformations at four locations: the Meu River at Montfort-sur-Meu (477 km^2): the forecasts are strongly biased; the Aa River at Wizernes (392 km^2): the uncertainty assessment is clearly under-dispersive; the Thérain River at Beauvais (755 km^2): the forecasts are reliable (except if no transformation is used) and the calibrated log-sinh and Box-Cox transformations (on $D2_{sup}$) are equivalent to the log transformation, which is here the best transformation on D3; the Sioulet River at Miremont (473 km^2): the calibration on $D2_{sup}$ leads to log-sinh and Box-Cox transformations equivalent to no transformation, which turns out not to be relevant on the control data set where the log and the Box-Cox transformations are more reliable.

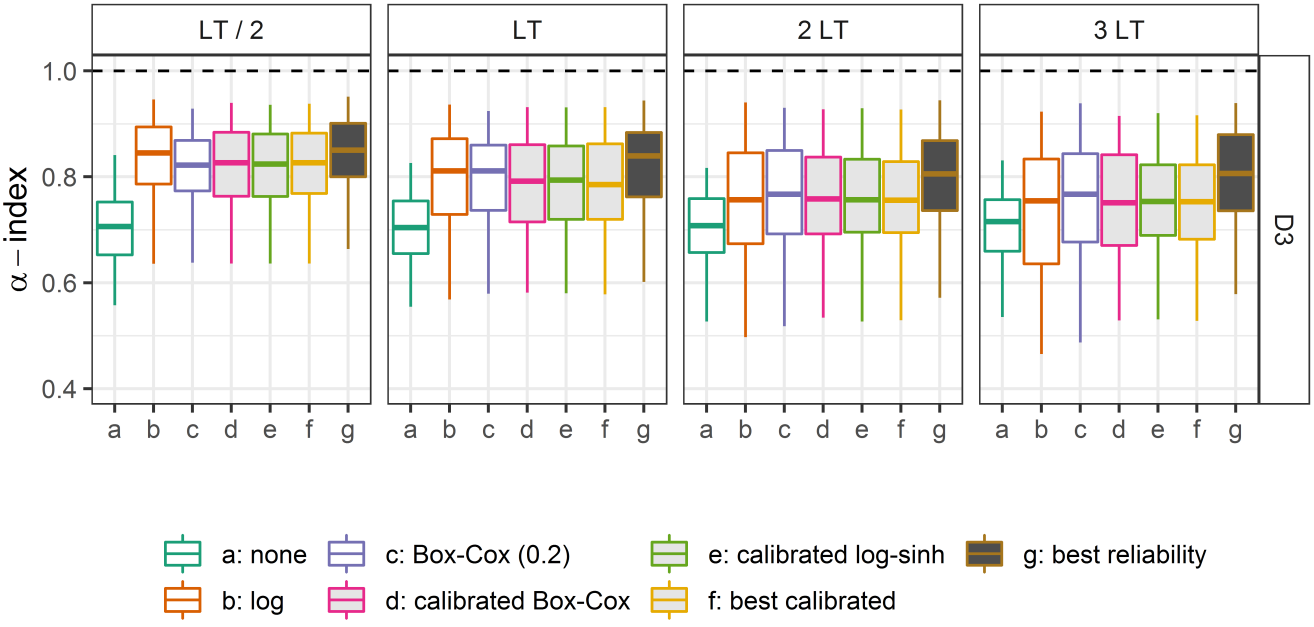

Figure 10. Distributions of the α-index values on the control data set D3, obtained with different transformations for four lead times (the filled boxplots represent the calibrated distributions). Boxplots (5th, 25th, 50th, 75th and 95th percentiles) synthesise the variety of scores over the catchments of the data set. Option "g" gives the best performance that could be achieved with this model and this post-processor for these catchments (see the discussion in Sect. 4.1).

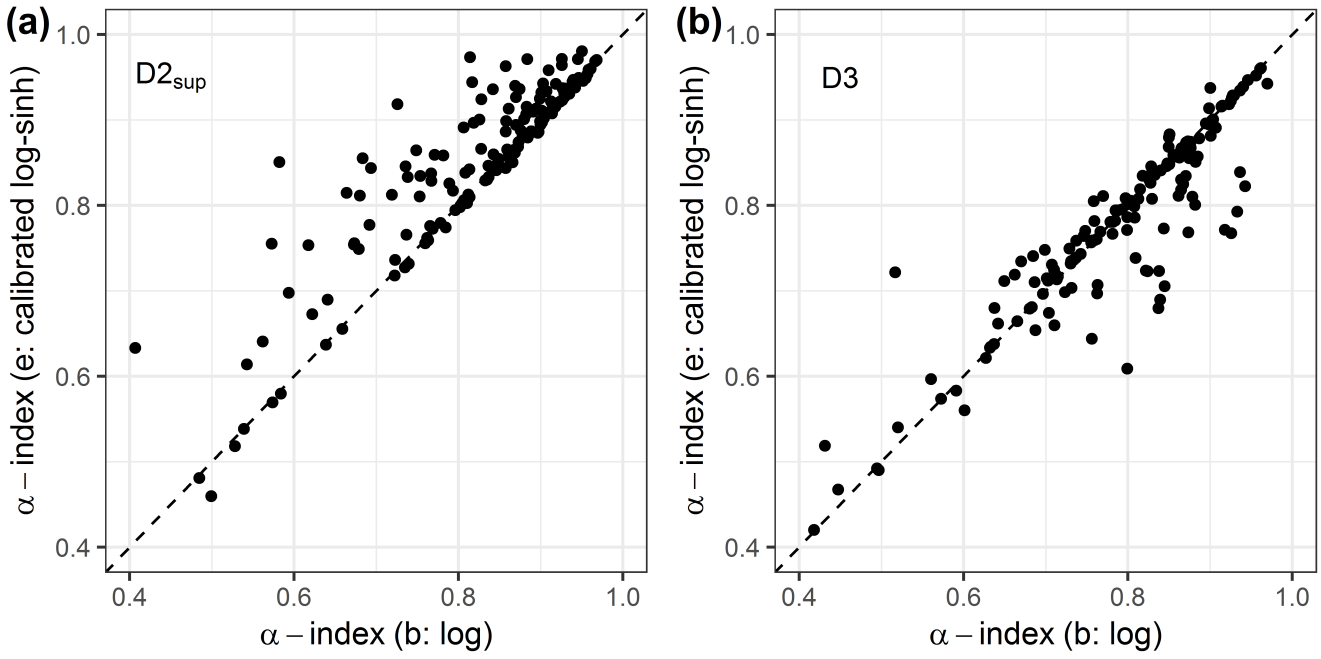

Figure 11. Scatter plots of the reliability α-index obtained with the log transformation and the log-sinh transformation: (a) on $D2_{sup}$ in the calibration step and (b) on D3 in the control step.

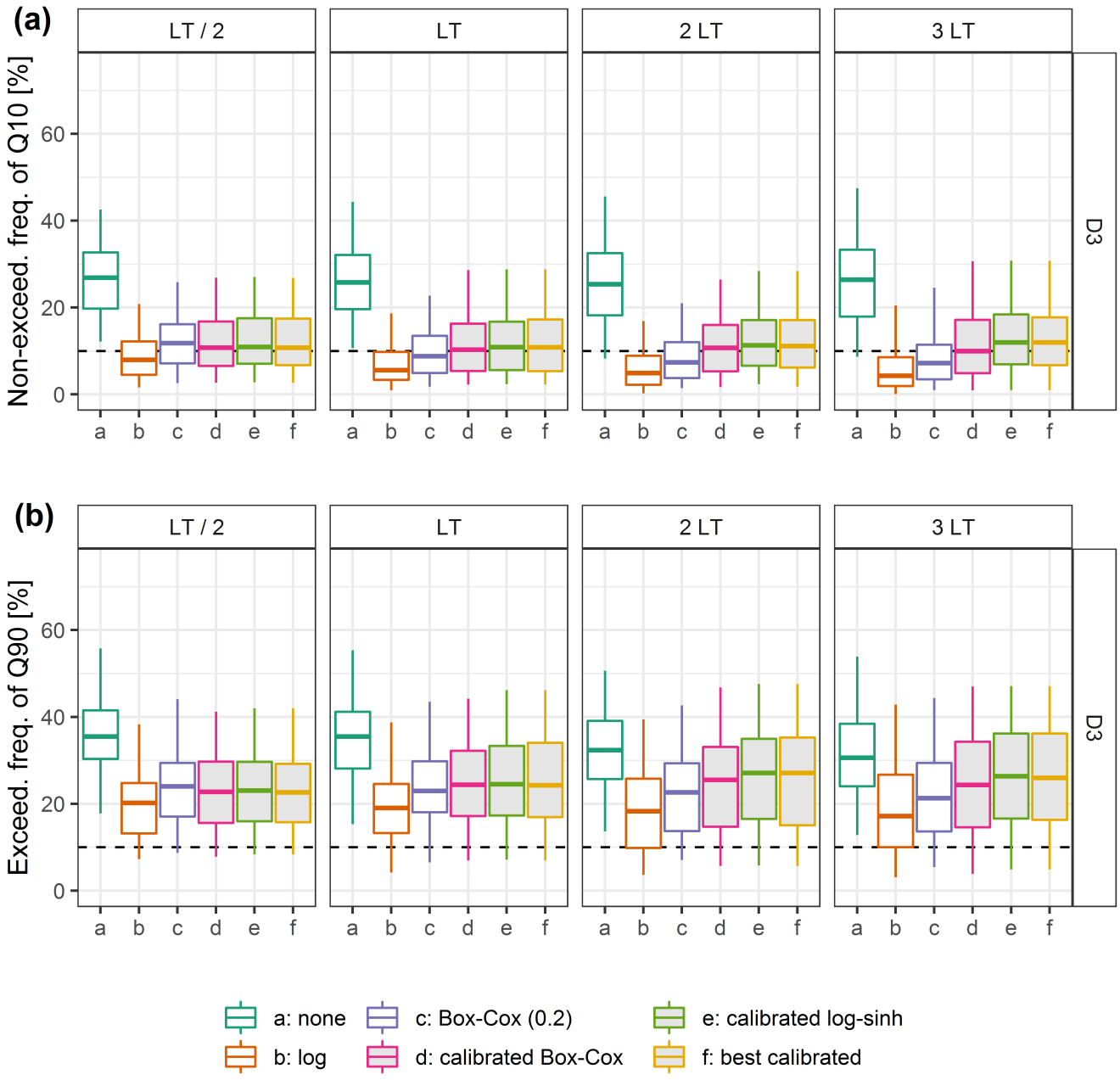

Figure 12. Distributions over the catchment set of a) the non-exceedance frequency of the 0.1 quantile and b) the exceedance frequency of the 0.9 quantile on the control data set D3, obtained with the different transformations tested (the filled boxplots are related to calibrated transformations).

~~Distributions over the catchment set of the non-exceedance frequency of the 0.1 quantile on the control data set D3, obtained with the different transformations tested (the filled boxplots are related to calibrated transformations).~~

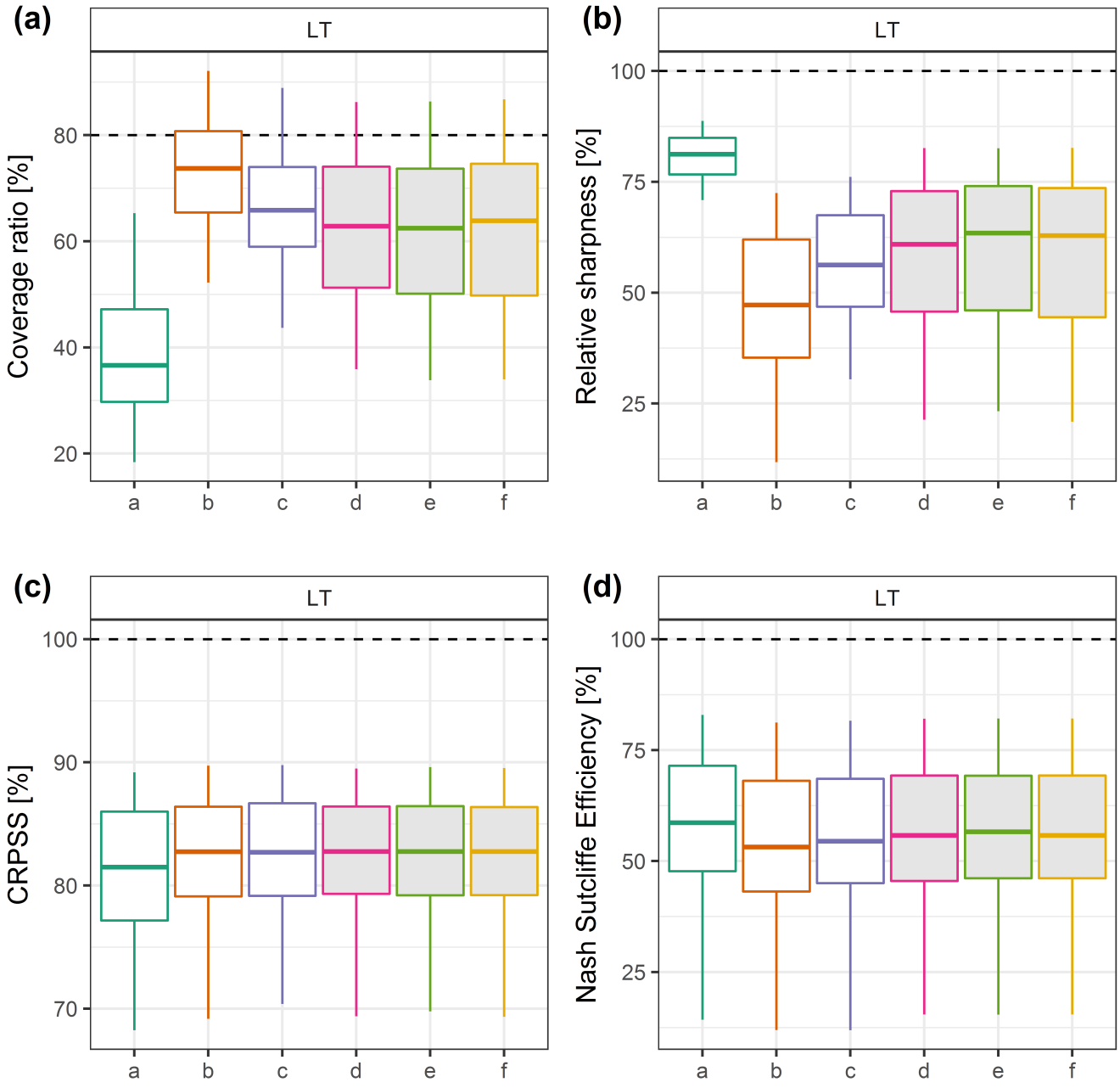

Figure 13. Distributions of coverage rate, relative sharpness, CRPSS and NSE values over the catchment set on the control data set D3, obtained with the different transformations tested (the filled boxplots are related to calibrated transformations).

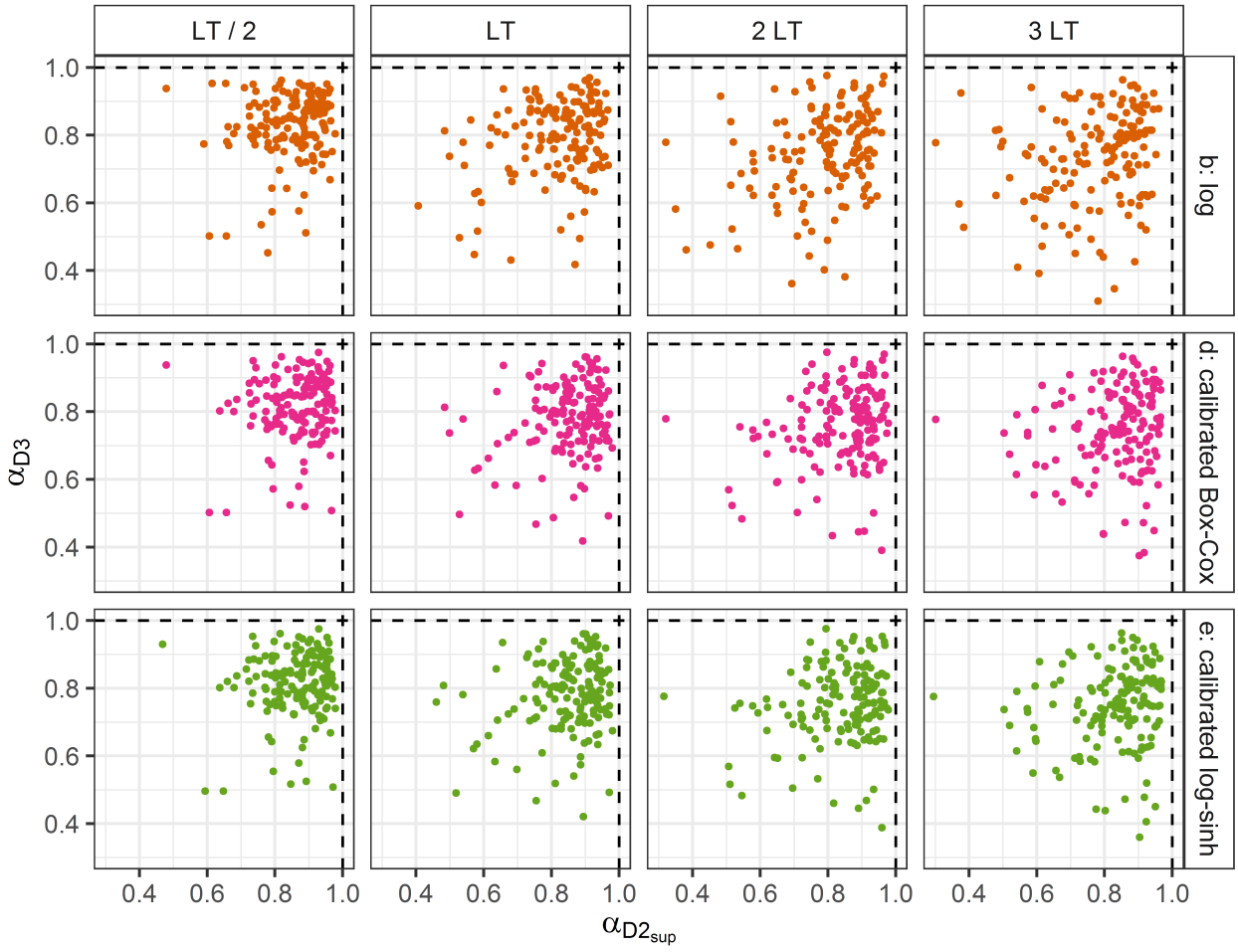

Figure 14. Comparison of the α-index values obtained in $D2_{sup}$ and D3. One point for each catchment. Similar results were obtained for the three other transformations (not shown).

4 Discussion

4.1 ~~Effect of the number of parameters of the~~ Do more complex parametric transformations yield better results in an extrapolation context?

Overall, the results obtained on the control data set suggest that the log transformation and the fixed Box-Cox transformation ($BC_{\lambda=0.2}$) can yield relatively satisfactory α-index and coverage ratio values given their multiplicative or near-multiplicative behaviour in extrapolation. More tapered behaviours that can be obtained with the calibrated Box-Cox or log-sinh transformations do not show advantages when extrapolating high flows on an independent data set. In other words, what is learnt during the calibration of the more complex parametric transformations does not yield better results in an extrapolation context.

These results could be explained by the fact that the calibration did not result in the optimally relevant parameter set. To investigate whether another calibration strategy could yield better results, we compared the performance on the D3 data set, when the calibration is achieved on $D2_{sup}$ ("f: best calibrated") or on D3 ("g: best reliability"). The results shown in Fig. ~~??~~ 10 reveal that even when the best parameter set is chosen among the 217 possibilities tested in this study (17 for the Box-Cox and 200 for the log-sinh), the α-index distributions are far from perfect and reliability decreases with increasing lead time. It suggests that the stability of the residuals distributions when extrapolating high flows might be a greater issue than the choice of the variable transformation. Nonetheless, the gap between the distributions of the non-parametric transformations ("b" and "c"), the best calibrated transformation ("f") and the best performance that could be achieved ("g") highlights that it might be possible to obtain better results with a more advanced calibration strategy. This is, however, beyond the scope of this study and is therefore left for further investigations.

~~Distributions of the α-index values on the control data set D3, obtained with different transformations for four lead times. Option "g" gives the best performance that could be achieved with this model and this post-processor for these catchments.~~

4.2 ~~Performance loss~~ Empirical-based versus distribution-based approaches : does the distribution shape choice impact the uncertainty assessment in an extrapolation context?

~~We investigated whether the reliability and reliability loss observed in an extrapolation context were correlated with any of the properties of the forecasts. First, Fig. 14 shows the relationship between the α-index values obtained in $D2_{sup}$ and those obtained in D3 for three representative transformations. The results indicate that it is not possible to anticipate the α-index values when extrapolating high flows in D3~~ Besides the reduction of heteroscedasticity, many studies use post-processors which are explicitly based on the ~~α-index values obtained when extrapolating high flows in $D2_{sup}$.~~

~~Comparison of the α-index values obtained in $D2_{sup}$ and D3. One point for each catchment. Similar results were obtained for the three other transformations (not shown).~~

~~In addition, two indices were chosen to describe the degree of extrapolation: the ratio of the median of the forecasted discharges on D3 over the median of the forecasted discharge on $D2_{sup}$ (Fig. **??**), and the ratio of the median of the forecasted discharges on D3 over the discharge for a return period of 20 years (rarity of the events of D3, for catchments where the~~

assessment of the vicennial discharge was available in the national database: http://www.hydro.eaufrance.fr, not shown). In both cases, no trend appears, regardless of the variable transformation used, with Spearman correlation coefficients lower than 0.5. The reliability can remain high for some catchments even when the magnitude of the events of the control data set are much higher than those of the training data set.

Reliability loss as a function of the rarity of the D3 events (ratio of the mean forecasted discharge D3 to mean forecasted discharge $D2_{sup}$ over the catchment set. Similar results were obtained for the three other transformations (not shown).

Finally, Fig. **??** shows the reliability loss as a function of the relative accuracy of the deterministic forecasts on $D2_{sup}$. A normalised RMSE was used to facilitate the visual representation, as in Lobligeois et al. (2014). Again, no clear trend is seen, which means that the goodness-of-fit during the calibration phase cannot be used as an indicator of the robustness of the uncertainty estimation in an extrapolation context.

Reliability loss as a function of the relative accuracy of the deterministic forecasts on $D2_{sup}$. Similar results were obtained for the three other transformations (not shown).

4.3 Empirical-based vs. distribution-based uncertainty assessment

Besides the reduction of heteroscedasticity, variable transformations may be used to fulfil the assumption of normality. Some post-processors are based on this hypothesis , such as assumption of a Gaussian distribution and use data transformations to fulfil this hypothesis (Li et al., 2017). Examples are the MCP or the meta-Gaussian model, and ; the NQT was designed to precisely achieve this. Here, we checked the normality it. In their study on autoregressive error models used as post-processors, Morawietz et al. (2011) showed that error models with an empirical distribution for the description of the standardized residuals perform better than those with a normal distribution. We first checked whether the variable transformation helped to reach a Gaussian distribution of the residuals computed with the transformed variablesusing . Then we investigated whether better performance can be achieved using empirical transformed residuals distributions or using Gaussian distributions calibrated on these empirical distributions.

We used the Shapiro-Francia test. For each parametric transformation, we selected the parameter set of the calibration grid which obtains the highest p-value. For more than 98% of the catchments, the p-value is lower than 0.018 (respectively, 0.023) when the Box-Cox transformation (respectively, the log-sinh transformation) is used. This indicates that there are only a few catchments for which the normality assumption is not to be rejected. In a nutshell, the variable transformations can stabilise the variance, but they do not necessarily ensure the normality of the residuals. It is important not to overlook this frequently encountered issue in hydrological studies.

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. For each transformation, the predictive uncertainty assessment obtained with the empirical transformed-variable residuals distribution is compared to the assessment based on the Gaussian distribution whose mean and variance are those of the empirical distribution. Figure 15 shows the α-index distributions obtained over the catchments for both options on the control data set D3. We note that no clear conclusion can be drawn. No transformation (or identity transformation), which does not reduce the het-

eroscedasticity at all, benefits from the use of the Gaussian distributions for all lead times. In contrast, the predictive uncertainty assessment based on the empirical distribution with the log transformation is more reliable than the one based on the Gaussian distribution. For short lead times, it is slightly better to use the empirical distributions for the calibrated transformations (Box-Cox and log-sinh), but we observe a different behaviour for longer lead times. For these longer lead times, assessing the predictive uncertainty by the Gaussian distribution fitted on the empirical distributions of transformed residuals obtained with the calibrated log-sinh or Box-Cox transformations is the most reliable option. It is better than using the log transformation with the empirical distribution, but not very different from using the $BC_{\lambda=0.2}$ transformation.

~~To further investigate~~ Investigations on the impact of the choice between the empirical or the Gaussian distributions ~~, Figure ?? shows the scatter plots of the α-index values obtained with the empirical distribution and with the Gaussian distribution for a lead time equal to the lag time. The values obtained for all catchments are well distributed on both sides of the bisector for the two transformations shown here and for the others (see the figures in Supplementary Material). There is no systematic behaviour: for some transformations, it is slightly better to choose the theoretical Gaussian quantiles, while empirical quantiles provide slightly more reliable predictive uncertainty assessment for others. It is important to note~~ on the post processor performance are shown in Supplementary materials. They show that the choice of the distribution is not the dominant factor~~: the variability of the distance to the bisector is much lower than the α-index variability obtained among catchments. The same pattern is observed for the other lead times that are not shown here (see the figures in Supplementary Material)~~.

~~Scatter plots of the α-index values obtained over the catchment set for two transformations, for the lead time LT: x-axis, using the empirical residuals distribution observed on the training data set; y-axis: using the Gaussian distribution fitted on the previous one.~~

4.3 ~~Links to previous results~~

~~As McInerney et al. (2017) pointed out, the choice of the heteroscedasticity modelling has a great impact on the predictive performance of probabilistic forecasts. Our results show that in an extrapolation context the log transformation provides overall the best results on the control data set for most catchments. The calibrated transformations are close to the log transformation for a significant number of catchments. Interestingly, for some catchments, the calibrated transformations are close to the no transformation on $D2_{sup}$, whereas the best choice on the control data set would have been another transformation. The $BC_{\lambda=0.2}$ transformation also yields reasonable results for a significant number of catchments. Nevertheless, for some sites, the best transformation is close to no transformation.~~

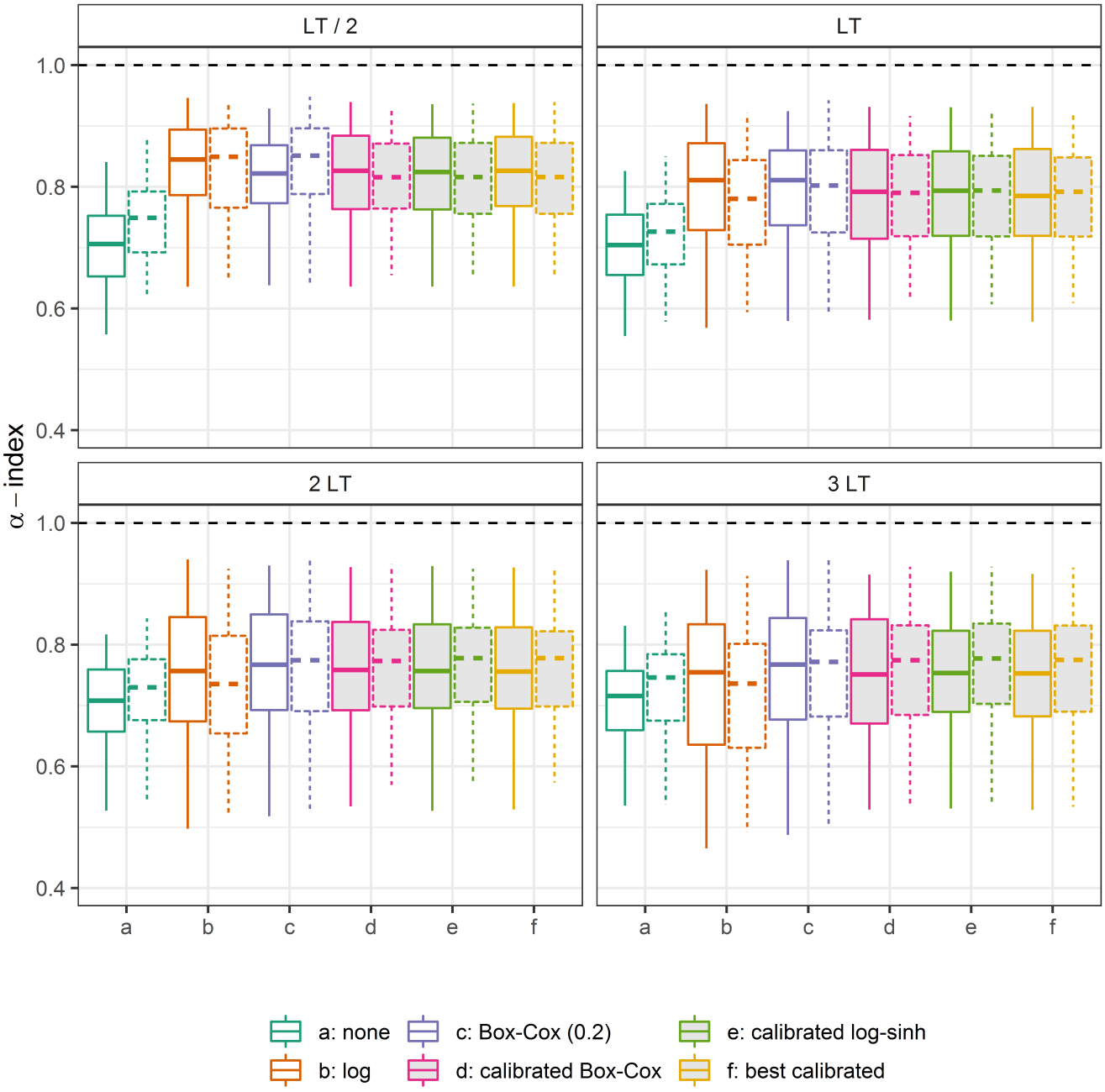

Figure 15. Distributions of the α-index values on the control data set D3, obtained with different transformations for four lead times, when using the empirical residuals distributions (straight boxplots) and the Gaussian distributions (dashed boxplots).

5 ~~A need for a focus change?~~

4.1 A need for a focus change?

In most modelling studies, several methodological steps depend on the range of the observations. First, calibration is designed to limit the residual errors in the available historical data. However the largest residuals are often associated with the highest discharge values. It is well known that removing the largest flood events from a data set can significantly modify the resulting calibrated parameter set. This is particularly true with the use of some common criteria such as quadratic criteria which strongly emphasise the largest errors (Legates and McCabe Jr., 1999; Berthet et al., 2010; Wright et al., 2015). Conversely, it is likely that "unavailable data" such as a physically realistic but (so far) unseen flood would significantly change the calibration results if it could be included in the calibration data set. Moreover, model conceptualisation (building) itself is often based on the understanding of how a catchment behaves "on average". In some studies, outliers may even be considered as disturbing and be discarded (Liano, 1996).

However, to provide robust models for operational purposes, we also need to focus on rare (rarely observed) events, still keeping in mind all the well-known issues associated with working with (too) few data (Anctil et al., 2004; Perrin et al., 2007). For predictive uncertainty assessment, this issue is exacerbated by the seasonality of hydrological extremes (Allamano et al., 2011; Li et al., 2013) for most approaches, which rely heavily on data (beyond data-learning approaches, all models which need to be calibrated). Therefore, there is an urgent need to gather and compile data on extreme events (Gaume et al., 2009). Nevertheless, operational forecasters must still prepare themselves to work in an extrapolation context, as pointed out by Andréassian et al. (2012).

5 Conclusions

Even if major floods are rare, it is of the utmost importance that the forecasts issued during such events are reliable to facilitate an efficient crisis management. Like Lieutenant Drogo in the Tartar Steppe who spent his entire life fulfilling his day-to-day duties, but waiting in his fortress for the invasion by foes (Buzzati, 1940), many forecasters expect and are preparing for a major event, even if their routine involves only minor events. That is why a strong concern for the extrapolation context should be encouraged in all modelling and calibration steps. This article proposes a control framework focusing on the forecasting performance in an extrapolation context.

We use this framework to test the predictive uncertainty assessment using a statistical post-processing of a rainfall-runoff model, based on a variable transformation. The latter has to handle the heteroscedasticity ~~of the predictive uncertainty~~and the evolution of the other predictive uncertainty distribution properties with the discharge magnitude, which is very problematic in an extrapolation context to issue reliable uncertainty assessment. As pointed out by McInerney et al. (2017), the choice of the heteroscedastic error modelling approach makes a significant impact on the predictive performance. This is true as well in an extrapolation context. ~~The main finding is thatthe~~

5.1 Main findings

Using the proposed framework for an evaluation in an extrapolation context, we showed that:

(a) Using an appropriate variable transformation can significantly improve the predictive distribution and its reliability. However, a performance loss still remains in an extrapolation context with any of the three transformations we tested.

(b) The more parametric transformations do not achieve significantly better results than the non-parametric transformations:

- while it allows a flexibility which can theoretically be very attractive in an extrapolation context, the log-sinh transformation is not more reliable in such a context;
- the non-parametric log transformation and Box-Cox transformation with the λ parameter set at 0.2 are robust and compare favourably.

(c) We did not find any variable significantly correlated with the performance loss in an extrapolation context.

The findings reported herein corroborate the results of McInerney et al. (2017) within the context of flood forecasting and extrapolation: calibrating the Box-Cox or log-sinh transformation can be counter-productive. We therefore suggest that operational flood forecasting services could consider the less flexible but more robust options: using the log transformation or the Box-Cox transformation with its λ parameter set ~~at~~ closed to 0.2~~or between 0.1 and 0.3~~.

Importantly, these results reveal significant performance losses on some catchments when it comes to extrapolation, whatever variable transformation is used. Even if the scheme tested yields satisfying results in terms of reliability for the majority of catchments, it fails on a significant number of catchments and further investigations are needed to gain a deeper understanding of when and why failures occur. ~~There are also some perspectives , for example with the~~

5.2 Limitations and perspectives

We used the framework designed by Krzysztofowicz et al. and already applied in various studies, which separates the input uncertainty (mainly the observed and forecasted rainfall) and the hydrological uncertainty. Our study focuses only on the 'effect' of the extrapolation degree in the hydrological uncertainty when using the best available rainfall product. Future works should combine both input uncertainty (rainfall) and hydrological uncertainty (Bourgin et al., 2014) to evaluate the impact of using uncertain (forecasted) rainfall in a forecasting context.

Though no variable was found to be correlated to the performance loss, the investigations should be continued using a wider set of variables. First, it may open new perspectives to explain these losses and improve our understanding of the flaws of the hydrological model and of the EHUP. Furthermore, it would be very useful to help operational forecasters to detect the hydrological situations for which their forecasts have to be questioned (in particular during major events when forecasts are made in an extrapolation context).

Furthermore, improving the calibration strategy and using a regionalisation of the predictive distribution assessment, as proposed in Bourgin et al. (2015) and Bock et al. (2018), ~~which~~ could help build more robust assessment of uncertainty quantification when forecasting high flows.

Finally, more studies focusing on the extrapolation context may help to better understand the limitations of the modelling (hydrological model structure, calibration, post-processing, etc.) and their consequences for practical matters. It is to be encouraged as a key for better and more reliable flood forecasting.

Appendix A: GRP model

The GRP model belongs to the suite of GR models (Michel, 1983). These models are designed to be as simple as possible but efficient for hydrological studies and for various operational needs, resulting in parsimonious model structures (https://webgr.irstea.fr/en/models/a-brief-history/, last access: 1 April 2019). The GRP model is designed for forecasting purposes (Berthet, 2010). It is a deterministic continuous lumped storage-type model (Fig. A1). The inputs are limited to areal rainfall and (interannual) potential evapotranspiration (both data may be available in real time). It can be understood as the sequence of two hydrological functions:

- a production function which is the same as in the well-known GR4J model developed by Perrin et al. (2003);
- a routing function which is a simplified version of the GR4J's routing function, since it only counts one flow branch composed of a unit hydrograph and quadratic routing store. The tests showed that the performance of the GRP and GR4J structures was similar in a forecasting mode.

A snow module (Valéry et al., 2014) may be implemented on top of the model if necessary.

Like any GR model, it is parsimonious. It has only three parameters: (a) an adjustment factor of effective rainfall which contributes to finding a good water balance, (b) the unit hydrogram time base used to account for the time lag between rainfall and streamflow and (c) the capacity of the routing store, which temporally smooths the output signal.

Its main difference with the other GR models is the implementation "in the loop" of two data assimilation schemes:

- a state updating procedure which modifies the main state of the routing function as a function of the last discharge values;
- an output updating based on an autoregressive model of the multiplicative error or an artificial neural network (multi-layer perception) whose inputs are the last discharge value and the two last forecasting errors. In this study, the autoregressive model was used.

The parameters are calibrated in forecasting mode, i.e., with the application of the updating procedures. This model is used by the French Flood Forecasting Services, some hydroelectricity suppliers and canal managers at the hourly time step in order to issue real-time forecasts for lead times ranging from a few hours to a few days at several hundred sites. Recently, Ficchì et al. (2016) pave the way to a multi-time step GRP version.

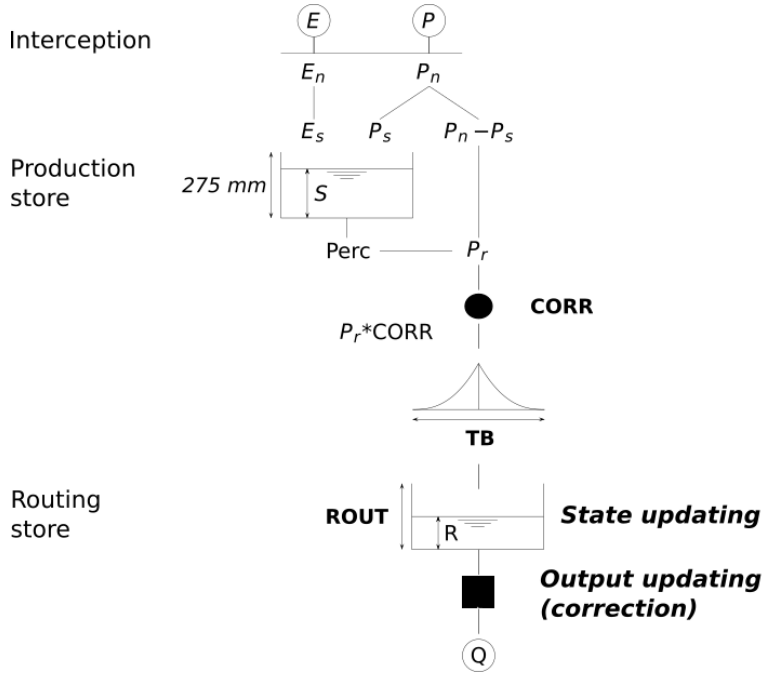

Figure A1. The GRP model flow-chart. After an interception step, the production function splits the net rainfall (P_n), according to the level of the production store. The effective rainfall (P_r) is the sum of the direct flow and the percolation from this store. A corrective multiplicative coefficient (CORR) is then applied. Then the flow runs through a unit hydrograph (time base: TB) and the routing store (capacity: ROUT).

Table B1. Summary of the different behaviours of the log-sinh transformation.

Cases	Behaviours
$\alpha > 3 \cdot \beta$	Additive error model
$y > 3 \cdot \beta$	Additive error model
$\alpha \ll y \ll \beta$	Multiplicative error model
$y \ll \alpha \ll \beta$	Additive error model
Otherwise if $\alpha + y \ll \beta$	Multiplicative error model (with an additive constant)

Appendix B: The log-sinh transformation behaviours

B1 Log-sinh transformation formulations

In this study, we used the formulations of the log-sinh transformation chosen by Del Giudice et al. (2013):

$$g_{\alpha,\beta 3\ [\alpha,\beta]} : y \mapsto \beta \cdot \log\left(\sinh\left(\frac{\alpha + y}{\beta}\right)\right) \tag{B1}$$

It is strictly equivalent to the formulation used by McInerney et al. (2017):

$$g'_{a,b 3\ [a,b]} : y \mapsto \frac{1}{b}\log\left(\sinh\left(a + b \cdot y\right)\right) \tag{B2}$$

B2 Different behaviours

Depending on the relative values of α and β and on the value range of y, compared to α and β, the log-sinh transformation can be reduced to an affine transformation (i.e. ~~$g_{\alpha,\beta}(y+\varepsilon) - g_{\alpha,\beta}(y) = \delta \cdot \varepsilon$~~ $g_{3\ [\alpha,\beta]}(y+\varepsilon) - g_{3\ [\alpha,\beta]}(y) = \delta \cdot \varepsilon$) or to the log transformation. The former case is equivalent to no transformation (identity function; additive error model), whereas the latter one is equivalent to a multiplicative error model (Table B1).

When $y \ll \beta$,

$$\begin{aligned}
\sinh\left(\frac{\alpha+y}{\beta}\right) &= \sinh\left(\frac{\alpha}{\beta}\right) \cdot \cosh\left(\frac{y}{\beta}\right) + \\
&\quad \sinh\left(\frac{y}{\beta}\right) \cdot \cosh\left(\frac{\alpha}{\beta}\right) \\
&\simeq \sinh\left(\frac{\alpha}{\beta}\right) \cdot \left(1 + \frac{y^2}{2 \cdot \beta^2} + \ldots\right) + \\
&\quad \cosh\left(\frac{\alpha}{\beta}\right) \cdot \left(\frac{y}{\beta} + \ldots\right)
\end{aligned}$$

Thus,

$$g_{\alpha,\beta 3\ [\alpha,\beta]}(y) \simeq \beta \cdot \log\left(\sinh\left(\frac{\alpha}{\beta}\right) + \cosh\left(\frac{\alpha}{\beta}\right) \cdot \frac{y}{\beta}\right)$$

When $\alpha \ll \beta$, the latter results in:

$$g_{\alpha,\beta 3\ [\alpha,\beta]}(y) \simeq \beta \cdot \log\left(\frac{\alpha+y}{\beta}\right) \tag{B3}$$

B2.1 Cases where the log-sinh transformation is equivalent to an affine transformation

If $x > 3$, $\sinh(x) \approx e^x/2$, then $\log(\sinh(x)) = x - \log(2)$. Therefore when $z = (\alpha+y)/\beta > 3$, the log-sinh transformation is equivalent to an affine transformation. In such cases, ~~$g^{-1}_{\alpha,\beta}(g_{\alpha,\beta}(y)+\varepsilon) = y+\varepsilon$~~ $g^{-1}_{3\ [\alpha,\beta]}(g_{3\ [\alpha,\beta]}(y)+\varepsilon) = y+\varepsilon$.

This happens when

- $\alpha > 3 \cdot \beta$;
- the β value is chosen large enough so that for any y value in the discharge range, $y > 3 \cdot \beta$.

Moreover, when $y \ll \alpha \ll \beta$, (B3) gives:

$$\begin{aligned} g_{\alpha,\beta 3\ [\alpha,\beta]} &\simeq \beta \cdot \left[\log\left(\frac{\alpha+y}{\alpha}\right) + \log(\alpha) - \log(\beta)\right] \\ &\simeq \frac{\beta \cdot y}{\alpha} + \beta \cdot (\log(\alpha) - \log(\beta)) \end{aligned}$$

The log-sinh transformation is then equivalent to an affine transformation.

B2.2 Cases where the log-sinh transformation is equivalent to a log transformation

~~When~~ As pointed out by McInerney et al. (2017), when $\alpha \ll y \ll \beta$, (B3) gives:

$$g_{\alpha,\beta 3\ [\alpha,\beta]}(y) \simeq \beta \cdot \log\left(\frac{y}{\beta}\right)$$

The log-sinh transformation is then equivalent to a mere log transformation.

B3 Calibration

The α and β parameters are in the same physical dimension as the y variable. Since this study is dedicated to the extrapolation context, we used the following adimensional ~~parametrization~~ parameterization to calibrate the variable transformation on various catchments. α and β are compared to the maximum forecasted discharge on the $D2_{sup}$ data subset:

$$\alpha = \gamma_1 \cdot \max_{D2}\left(\widetilde{Q}\right) \text{ and } \beta = \gamma_2 \cdot \max_{D2}\left(\widetilde{Q}\right) \tag{B4}$$

In the calibration step, the parameter space is explored on a (γ_1, γ_2) grid: 18 values of γ_1 from 0.01 to 100 and 15 values of γ_2 from 0.1 to 100 were tested, excluding combined values leading to the very same behaviours, such as $\gamma_1 \gg 3 \cdot \gamma_2$, equivalent to no transformation (additive error model). A total of 200 (γ_1, γ_2) combinations were tested for each calibration.

Appendix C: ~~Supplementary material~~

~~Distributions of coverage rate, relative sharpness, CRPSS and NSE values over the catchment set on control data set D3, obtained with the different transformations tested (the filled boxplots are related to calibrated transformations), for lead time LT / 2.~~

~~Distributions of coverage rate, relative sharpness, CRPSS and NSE values over the catchment set on control data set D3, obtained with the different transformations tested (the filled boxplots are related to calibrated transformations), for lead time 2 LT.~~

~~Distributions of coverage rate, relative sharpness, CRPSS and NSE values over the catchment set on control data set D3, obtained with the different transformations tested (the filled boxplots are related to calibrated transformations), for lead time 3 LT.~~

~~Scatter plots of the reliability α-index over the catchment set: x-axis, using the empirical residuals distribution observed on the training data set; y-axis: using the Gaussian distribution fitted on the previous one.~~

Author contributions. All co-authors contributed to and edited the paper. These authors contributed equally: Lionel Berthet, François Bourgin; they should be both considered as first authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors thank Dr. Kolbjorn Engeland and one anonymous reviewer. Both provided insightful comments which helped to greatly improve this text. Thanks are also due to Météo-France for providing the meteorological data and Banque HYDRO data base for the hydrological data. The contribution of the authors from Irstea was financially supported by SCHAPI (Ministry of the Environment).

References

Abaza, M., Anctil, F., Fortin, V., and Perreault, L.: On the incidence of meteorological and hydrological processors: Effect of resolution, sharpness and reliability of hydrological ensemble forecasts, Journal of Hydrology, 555, 371–384, https://doi.org/10.1016/j.jhydrol.2017.10.038, 2017.

Abbaszadeh, P., Moradkhani, H., and Yan, H.: Enhancing hydrologic data assimilation by evolutionary Particle Filter and Markov Chain Monte Carlo, Advances in Water Resources, 111, 192–204, https://doi.org/10.1016/j.advwatres.2017.11.011, 2018.

Allamano, P., Laio, F., and Claps, P.: Effects of disregarding seasonality on the distribution of hydrological extremes, Hydrology and Earth System Sciences, 15, 3207–3215, https://doi.org/10.5194/hess-15-3207-2011, 2011.

Anctil, F., Perrin, C., and Andréassian, V.: Impact of the length of observed records on the performance of ANN and of conceptual parsimonious rainfall-runoff forecasting models, Environmental Modelling & Software, 19, 357 – 368, 2004.

Andréassian, V., Hall, A., Chahinian, N., and Schaake, J.: Introduction and Synthesis: Why should hydrologists work on a large number of basin data sets?, IAHS-AISH Publication, n° 307, 1–5, 2006.

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M. H., and Valery, A.: HESS Opinions 'Crash tests for a standardized evaluation of hydrological models', Hydrology and Earth System Sciences, 13, 1757–1764, 2009.

Andréassian, V., Le Moine, N., Perrin, C., Ramos, M.-H., Oudin, L., Mathevet, T., Lerat, J., and Berthet, L.: All that glitters is not gold: the case of calibrating hydrological models, Hydrological Processes, 26, 2206–2210, https://doi.org/10.1002/hyp.9264, 2012.

Barbetta, S., Coccia, G., Moramarco, T., Brocca, L., and Todini, E.: The multi temporal/multi-model approach to predictive uncertainty assessment in real-time flood forecasting, Journal of Hydrology, 551, 555–576, https://doi.org/10.1016/j.jhydrol.2017.06.030, 2017.

Bennett, J. C., Robertson, D. E., Shrestha, D. L., Wang, Q., Enever, D., Hapuarachchi, P., and Tuteja, N. K.: A System for Continuous Hydrological Ensemble Forecasting (SCHEF) to lead times of 9 days, Journal of Hydrology, 519, Part D, 2832–2846, https://doi.org/10.1016/j.jhydrol.2014.08.010, 2014.

Berthet, L.: Flood forecasting at the hourly time-step: for a better assimilation of flow information in hydrological modelling (in French), Ph.D. thesis, AgroParisTech (Paris), Irstea (Antony), Doctoral School GRN, 2010.

Berthet, L. and Piotte, O.: International survey for good practices in forecasting uncertainty assessment and communication, in: EGU General Assembly, vol. 16, pp. EGU2014–8579, 2014.

Berthet, L., Andreassian, V., Perrin, C., and Javelle, P.: How crucial is it to account for the antecedent moisture conditions in flood forecasting? Comparison of event-based and continuous approaches on 178 catchments, Hydrology and Earth System Sciences, 13, 819–831, 2009.

Berthet, L., Andréassian, V., Perrin, C., and Loumagne, C.: How significant are quadratic criteria? Part 2. On the relative contribution of large flood events to the value of a quadratic criterion, Hydrological Sciences Journal, 55, 1063–1073, https://doi.org/10.1080/02626667.2010.505891, 2010.

Bock, A. R., Farmer, W. H., and Hay, L. E.: Quantifying uncertainty in simulated streamflow and runoff from a continental-scale monthly water balance model, Advances in Water Resources, 122, 166 – 175, https://doi.org/10.1016/j.advwatres.2018.10.005, 2018.

Bogner, K., Pappenberger, F., and Cloke, H. L.: Technical Note: The normal quantile transformation and its application in a flood forecasting system, Hydrology and Earth System Sciences, 16, 1085–1094, https://doi.org/10.5194/hess-16-1085-2012, 2012.

Bourgin, F.: How to quantify predictive uncertainty in hydrological modelling? Exploratory work on a large sample of catchments (in French), Ph.D. thesis, AgroParisTech (Paris), Irstea (Antony), Doctoral School GRNE, 2014.

Bourgin, F., Ramos, M., Thirel, G., and Andréassian, V.: Investigating the interactions between data assimilation and post-processing in hydrological ensemble forecasting, Journal of Hydrology, 519, 2775–2784, https://doi.org/10.1016/j.jhydrol.2014.07.054, 2014.

Bourgin, F., Andréassian, V., Perrin, C., and Oudin, L.: Transferring global uncertainty estimates from gauged to ungauged catchments, Hydrology and Earth System Sciences, 19, 2535–2546, https://doi.org/10.5194/hess-19-2535-2015, 2015.

Box, G. E. P. and Cox, D. R.: An Analysis of Transformations, Journal of the Royal Statistical Society. Series B (Methodological), 26, 211–252, 1964.

Breiman, L.: Statistical modeling: The two cultures, Statistical Sciences, 16, 199–215, https://doi.org/10.1214/ss/1009213726, 2001.

Bremnes, J. B.: Constrained Quantile Regression Splines for Ensemble Postprocessing, Monthly Weather Review, 147, 1769–1780, https://doi.org/10.1175/MWR-D-18-0420.1, https://doi.org/10.1175/MWR-D-18-0420.1, 2019.

Brigode, P., Oudin, L., and Perrin, C.: Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change?, Journal of Hydrology, 476, 410–425, https://doi.org/10.1016/j.jhydrol.2012.11.012, 2013.

Buzzati, D.: Il deserto dei Tartari (The Tartar Steppe), Rizzoli, Milano, 1940.

Cigizoglu, H.: Estimation, forecasting and extrapolation of river flows by artificial neural networks, Hydrological Sciences Journal, 48, 349–362, 2003.

Coccia, G. and Todini, E.: Recent developments in predictive uncertainty assessment based on the model conditional processor approach, Hydrology and Earth System Sciences, 15, 3253–3274, https://doi.org/10.5194/hess-15-3253-2011, 2011.

Coron, L., Andreassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., and Hendrickx, F.: Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments, Water Resources Research, 48, W05 552, https://doi.org/10.1029/2011wr011721, 2012.

Del Giudice, D., Honti, M., Scheidegger, A., Albert, C., Reichert, P., and Rieckermann, J.: Improving uncertainty estimation in urban hydrological modeling by statistically describing bias, Hydrology and Earth System Sciences, 17, 4209–4225, https://doi.org/10.5194/hess-17-4209-2013, 2013.

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H. D., Fresch, M., Schaake, J., and Zhu, Y.: The Science of NOAA's Operational Hydrologic Ensemble Forecast Service, Bulletin of the American Meteorological Society, 95, 79–98, https://doi.org/10.1175/BAMS-D-12-00081.1, 2014.

Demeritt, D., Cloke, H., Pappenberger, F., Thielen, J., Bartholmes, J., and Ramos, M.-H.: Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting, Environmental Hazards, 7, 115–127, 2007.

Dogulu, N., López López, P., Solomatine, D. P., Weerts, A. H., and Shrestha, D. L.: Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods and their comparison on contrasting catchments, Hydrology and Earth System Sciences, 19, 3181–3201, https://doi.org/10.5194/hess-19-3181-2015, 2015.

Ficchì, A., Perrin, C., and Andréassian, V.: Impact of temporal resolution of inputs on hydrological model performance: An analysis based on 2400 flood events, Journal of Hydrology, 538, 454–470, https://doi.org/10.1016/j.jhydrol.2016.04.016, 2016.

Furusho, C., Perrin, C., Viatgé, J., R., L., and Andréassian, V.: Collaborative work between operational forecasters and scientists for better flood forecasts, La Houille Blanche, pp. 5–10 (in French), https://doi.org/10.1051/lhb/2016033, 2016.

Gaume, E., Bain, V., Bernardara, P., Newinger, O., Barbuc, M., Bateman, A., Blaskovicová, L., Blöschl, G., Borga, M., Dumitrescu, A., Daliakopoulos, I., Garcia, J., Irimescu, A., Kohnova, S., Koutroulis, A., Marchi, L., Matreata, S., Medina, V., Preciso, E., Sempere-Torres, D., Stancalie, G., Szolgay, J., Tsanis, I., Velasco, D., and Viglione, A.: A compilation of data on European flash floods, Journal of Hydrology, 367, 70 – 78, https://doi.org/10.1016/j.jhydrol.2008.12.028, 2009.

Giustolisi, O. and Laucelli, D.: Improving generalization of artificial neural networks in rainfall-runoff modelling, Hydrological Sciences Journal, 50, 439–457, 2005.

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, https://doi.org/10.1111/j.1467-9868.2007.00587.x, 2007.

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., and Andréassian, V.: Large-sample hydrology: a need to balance depth with breadth, Hydrology and Earth System Sciences, 18, 463–477, https://doi.org/10.5194/hess-18-463-2014, 2014.

Hemri, S., Lisniak, D., and Klein, B.: Multivariate postprocessing techniques for probabilistic hydrological forecasting, Water Resources Research, 51, 7436–7451, https://doi.org/10.1002/2014WR016473, 2015.

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather and Forecasting, 15, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:dotcrp>2.0.co;2, 2000.

Imrie, C., Durucan, S., and Korre, A.: River flow prediction using artificial neural networks: Generalisation beyond the calibration range, Journal of Hydrology, 233, 138–153, 2000.

Klemeš, V.: Operational testing of hydrological simulation models, Hydrological Sciences Journal-Journal Des Sciences Hydrologiques, 31, 13–24, https://doi.org/10.1080/02626668609491024, 1986.

Krzysztofowicz, R.: Bayesian theory of probabilistic forecasting via deterministic hydrologic model, Water Resources Research, 35, 2739–2750, 1999.

Krzysztofowicz, R. and Maranzano, C. J.: Hydrologic uncertainty processor for probabilistic stage transition forecasting, Journal of Hydrology, 293, 57–73, https://doi.org/10.1016/j.jhydrol.2004.01.003, 2004.

Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrology and Earth System Sciences, 11, 1267–1277, 2007.

Lang, M., Pobanz, K., Renard, B., Renouf, E., and Sauquet, E.: Extrapolation of rating curves by hydraulic modelling, with application to flood frequency analysis, Hydrological Sciences Journal, 55, 883–898, https://doi.org/10.1080/02626667.2010.504186, 2010.

Legates, D. and McCabe Jr., G.: Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation, Water Resources Research, 35, 233–241, 1999.

Li, M., Wang, Q. J., and Bennett, J.: Accounting for seasonal dependence in hydrological model errors and prediction uncertainty, Water Resources Research, 49, 5913–5929, https://doi.org/10.1002/wrcr.20445, 2013.

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., and Di, Z.: A review on statistical postprocessing methods for hydrometeorological ensemble forecasting, Wiley Interdisciplinary Reviews: Water, 4, e1246, https://doi.org/10.1002/wat2.1246, 2017.

Liano, K.: Robust error measure for supervised neural network learning with outliers, IEEE Transactions on Neural Networks, 7, 246–250, 1996.

Lobligeois, F., Andréassian, V., Perrin, C., Tabary, P., and Loumagne, C.: When does higher spatial resolution rainfall information improve streamflow simulation? An evaluation using 3620 flood events, Hydrology and Earth System Sciences, 18, 575–594, https://doi.org/10.5194/hess-18-575-2014, 2014.

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., and Kuczera, G.: Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors, Water Resources Research, 53, 2199–2239, https://doi.org/10.1002/2016WR019168, 2017.

Merz, R., Parajka, J., and Blöschl, G.: Time stability of catchment model parameters: Implications for climate impact analyses, Water Resources Research, 47, W02 531, https://doi.org/10.1029/2010WR009505, 2011.

Michel, C.: Que peut-on faire en hydrologie avec un modèle conceptuel à un seul paramètre?, La Houille blanche, pp. 39–44 (in French), 1983.

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

Montanari, A. and Grossi, G.: Estimating the uncertainty of hydrological forecasts: A statistical approach, Water Resources Research, 44, W00B08, https://doi.org/10.1029/2008wr006897, 2008.

Moradkhani, H., Hsu, K. L., Gupta, H., and Sorooshian, S.: Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, Water Resources Research, 41, W05 012, https://doi.org/10.1029/2004wr003604, 2005a.

Moradkhani, H., Sorooshian, S., Gupta, H. V., and Houser, P. R.: Dual state–parameter estimation of hydrological models using ensemble Kalman filter, Advances in Water Resources, 28, 135–147, https://doi.org/10.1016/j.advwatres.2004.09.002, 2005b.

Morawietz, M., Xu, C.-Y., Gottschalk, L., and Tallaksen, L. M.: Systematic evaluation of autoregressive error models as post-processors for a probabilistic streamflow forecast system, Journal of Hydrology, 407, 58–72, https://doi.org/10.1016/j.jhydrol.2011.07.007, 2011.

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, https://doi.org/10.1016/j.jhydrol.2004.08.026, 2005.

Pagano, T. C., Shrestha, D. L., Wang, Q. J., Robertson, D., and Hapuarachchi, P.: Ensemble dressing for hydrological applications, Hydrological Processes, 27, 106–116, https://doi.org/10.1002/hyp.9313, 2013.

Pagano, T. C., Wood, A. W., Ramos, M.-H., Cloke, H. L., Pappenberger, F., Clark, M. P., Cranston, M., Kavetski, D., Mathevet, T., Sorooshian, S., and Verkade, J. S.: Challenges of Operational River Forecasting, J. Hydrometeor., 15, 1692–1707, https://doi.org/10.1175/JHM-D-13-0188.1, 2014.

Pappenberger, F. and Beven, K. J.: Ignorance is bliss: Or seven reasons not to use uncertainty analysis, Water Resources Research, 42, W05 302, https://doi.org/10.1029/2005wr004820, 2006.

Pappenberger, F., Thielen, J., and Del Medico, M.: The impact of weather forecast improvements on large scale hydrology: analysing a decade of forecasts of the European Flood Alert System, Hydrological Processes, 25, 1091–1113, https://doi.org/10.1002/hyp.7772, 2011.

Pappenberger, F., Pagano, T. C., Brown, J. D., Alfieri, L., Lavers, D. A., Berthet, L., Bressand, F., Cloke, H. L., Cranston, M., Danhelka, J., Demargne, J., Demuth, N., de Saint-Aubin, C., Feikema, P. M., Fresch, M. A., Garçon, R., Gelfan, A., He, Y., Hu, Y. Z., Janet, B., Jurdy, N., Javelle, P., Kuchment, L., Laborda, Y., Langsholt, E., Le Lay, M., Li, Z. J., Mannessiez, F., Marchandise, A., Marty, R., Meißner, D., Manful, D., Organde, D., Pourret, V., Rademacher, S., Ramos, M. H., Reinbold, D., Tibaldi, S., Silvano, P., Salamon, P., Shin, D., Sorbet, C., Sprokkereef, E., Thiemig, V., Tuteja, N. K., van Andel, S. J., Verkade, J. S., Vehviläinen, B., Vogelbacher, A., Wetterhall, F., Zappa, M., Van der Zwan, R. E., and Thielen-del Pozo, J.: Hydrological Ensemble Prediction Systems Around the Globe, pp. 1–35, Springer Berlin Heidelberg, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-40457-3_47-1, 2016.

Perrin, C., Oudin, L., Andréassian, V., Rojas-Serna, C., Michel, C., and Mathevet, T.: Impact of limited streamflow data on the efficiency and the parameters of rainfall-runoff models, Hydrological Sciences Journal, 52, 131–151, 2007.

Ramos, M.-H., Bartholmes, J., and Pozo, J. T.-d.: Development of decision support products based on ensemble forecasts in the European flood alert system, Atmospheric Science Letters, 8, 113–119, https://doi.org/10.1002/asl.161, 2007.

Reichert, P. and Mieleitner, J.: Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters, Water Resources Research, 45, W10 402, https://doi.org/10.1029/2009wr007814, 2009.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resources Research, 46, W05 521, https://doi.org/10.1029/2009wr008328, 2010.

Salamon, P. and Feyen, L.: Assessing parameter, precipitation, and predictive uncertainty in a distributed hydrological model using sequential data assimilation with the particle filter, Journal of Hydrology, 376, 428–442, https://doi.org/10.1016/j.jhydrol.2009.07.051, 2009.

Schoups, G. and Vrugt, J. A.: A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, Water Resources Research, 46, W10 531, https://doi.org/10.1029/2009wr008933, 2010.

Seiller, G., Anctil, F., and Roy, R.: Design and experimentation of an empirical multistructure framework for accurate, sharp and reliable hydrological ensembles, Journal of Hydrology, 552, 313–340, https://doi.org/10.1016/j.jhydrol.2017.07.002, 2017.

Sharma, S., Siddique, R., Reed, S., Ahnert, P., Mendoza, P., and Mejia, A.: Relative effects of statistical preprocessing and postprocessing on a regional hydrological ensemble prediction system, Hydrology and Earth System Sciences, 22, 1831–1849, https://doi.org/10.5194/hess-22-1831-2018, 2018.

Singh, S. K., McMillan, H., and Bardossy, A.: Use of the data depth function to differentiate between case of interpolation and extrapolation in hydrological model prediction, Journal of Hydrology, 477, 213–228, https://doi.org/10.1016/j.jhydrol.2012.11.034, 2013.

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

Tabary, P., Dupuy, P., L'Henaff, G., Gueguen, C., Moulin, L., Laurantin, O., Merlier, C., and Soubeyroux, J.-M.: A 10-year (1997-2006) reanalysis of Quantitative Precipitation Estimation over France: methodology and first results, vol. 351, pp. 255–260, IAHS, 2012.

Thielen, J., Bartholmes, J., Ramos, M.-H., and de Roo, A.: The European Flood Alert System - Part 1: Concept and development, Hydrology and Earth System Sciences, 13, 125 – 140, 2009.

Todini, E.: Role and treatment of uncertainty in real-time flood forecasting, Hydrological Processes, 18, 2743–2746, https://doi.org/10.1002/hyp.5687, 2004.

Todini, E.: Hydrological catchment modelling: past, present and future, Hydrology and Earth System Sciences, 11, 468–482, 2007.

Todini, E.: A model conditional processor to assess predictive uncertainty in flood forecasting, International Journal of River Basin Management, 6, 123–137, https://doi.org/10.1080/15715124.2008.9635342, 2008.

Todini, E.: Predictive uncertainty assessment in real time flood forecasting, in: Uncertainties in Environmental Modelling and Consequences for Policy Making, edited by Baveye, P. C., Laba, M., and Mysiak, J., pp. 205–228, Springer Netherlands, Dordrecht, 2009.

Valéry, A., Andréassian, V., and Perrin, C.: 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2 – Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments, Journal of Hydrology, 517, 1176 – 1187, https://doi.org/10.1016/j.jhydrol.2014.04.058, 2014.

Van Steenbergen, N., Ronsyn, J., and Willems, P.: A non-parametric data-based approach for probabilistic flood forecasting in support of uncertainty communication, Environmental Modelling & Software, 33, 92–105, https://doi.org/10.1016/j.envsoft.2012.01.013, 2012.

Vaze, J., Post, D., Chiew, F., Perraud, J.-M., Viney, N., and Teng, J.: Climate non-stationarity – Validity of calibrated rainfall–runoff models for use in climate change studies, Journal of Hydrology, 394, 447–457, https://doi.org/10.1016/j.jhydrol.2010.09.018, 2010.

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, https://doi.org/10.5194/hess-14-2303-2010, 2010.

Verkade, J., Brown, J., Davids, F., Reggiani, P., and Weerts, A.: Estimating predictive hydrological uncertainty by dressing deterministic and ensemble forecasts; a comparison, with application to Meuse and Rhine, Journal of Hydrology, 555, 257–277, https://doi.org/10.1016/j.jhydrol.2017.10.024, 2017.

Verkade, J. S. and Werner, M. G. F.: Estimating the benefits of single value and probability forecasting for flood warning, Hydrology and Earth System Sciences, 15, 3751–3765, https://doi.org/10.5194/hess-15-3751-2011, 2011.

Viatgé, J., Pinna, T., Perrin, C., Dorchies, D., and Garandeau, L.: Towards an enhanced temporal flexibility of the GRP flood forecasting operational model, in: Proceedings of the SHF conference *De la prévision des crues à la gestion de crise* (From flood forecasting to crisis management), Avignon (France), November 14th - 16th, Société Hydrotechnique de France, 12 p. (in French), 2018.

Viatgé, J., Berthet, L., Marty, R., Bourgin, F., Piotte, O., , Ramos, M. H., and Perrin, C.: Towards the real-time production of predictive intervals around streamflow forecasts in Vigicrues in France, LHB, pp. 63–71, https://doi.org/10.1051/lhb/2019016, 2019.

Wang, Q. J., Robertson, D. E., and Chiew, F. H. S.: A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites, Water Resources Research, 45, W05 407, https://doi.org/10.1029/2008wr007355, 2009.

Wang, Q. J., Shrestha, D. L., Robertson, D. E., and Pokhrel, P.: A log-sinh transformation for data normalization and variance stabilization, Water Resources Research, 48, W05 514, https://doi.org/10.1029/2011wr010973, 2012.

Wani, O., Beckers, J. V. L., Weerts, A. H., and Solomatine, D. P.: Residual uncertainty estimation using instance-based learning with applications to hydrologic forecasting, Hydrology and Earth System Sciences, 21, 4021–4036, https://doi.org/10.5194/hess-21-4021-2017, 2017.

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, https://doi.org/10.5194/hess-15-255-2011, 2011.

Wilby, R. L.: Uncertainty in water resource model parameters used for climate change impact assessment, Hydrological Processes, 19, 3201–3219, https://doi.org/10.1002/hyp.5819, 2005.

Woldemeskel, F., McInerney, D., Lerat, J., Thyer, M., Kavetski, D., D, S., Tuteja, N., and Kuczera, G.: Evaluating post-processing approaches for monthly and seasonal streamflow forecasts, Hydrology and Earth System Sciences, 22, 6257–6278, https://doi.org/10.5194/hess-22-6257-2018, 2018.

Wright, D. P., Thyer, M., and Westra, S.: Influential point detection diagnostics in the context of hydrological model calibration, Journal of Hydrology, 527, 1161 – 1172, https://doi.org/10.1016/j.jhydrol.2015.05.047, 2015.

Yang, J., Reichert, P., and Abbaspour, K. C.: Bayesian uncertainty analysis in distributed hydrologic modeling: a case study in the Thur River basin (Switzerland), Water Resources Research, 43, W10 401, https://doi.org/10.1029/2006wr005497, 2007a.

Yang, J., Reichert, P., Abbaspour, K. C., and Yang, H.: Hydrological modelling of the Chaohe Basin in China: Statistical model formulation and Bayesian inference, Journal of Hydrology, 340, 167–182, https://doi.org/10.1016/j.jhydrol.2007.03.006, 2007b.

Zalachori, I., Ramos, M. H., Garçon, R., Mathevet, T., and Gailhard, J.: Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies, Advances in Sciences and Research, 8, 135–141, https://doi.org/10.5194/asr-8-135-2012, 2012.