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

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.

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.
- It is not clear what are the inputs and outputs of the EHUP.
- Why does the EHUP require a separate training period to transformation calibration period?

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**
  - **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.
- **Section 4.2**
  - This sub-section seems like it is attempting to determine what the key drivers for performance are, but this is not evident from heading.
  - Since the authors did not find any key drivers for poor performance, I'm not sure if this analysis adds much value.
- **Section 4.3**
  - 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.

    - 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.
- **Section 4.4**
    - 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.
- **Section 5:** “A need for a focus change.”
    - This section is not long enough for a separate section, and is a discussion topic. I suggest moving this to the discussion.
- **Limitations and future work**
    - It would be useful to have a sub-section discussing the limitations of this study and future research.

Too many figures. I believe this paper has too many figures. I recommend

- Merging some figures
    - fig 6 and 10
    - fig 12 and 13
- Is there any point in showing all 3 transformations for fig 16-18?
    - You could consider a single transformation and combine into a single figure.
    - Or you could move fig 17-18 to sup mat since they don’t show any correlations.

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).

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?

Table 1: Change “percentiles” to “quantiles”

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.

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?

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.

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?

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.

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

Pg 10, line 6: “maximum discharge of time series”

Make it clear you are referring to *forecast* discharge here.

Pg 10, line 8: “the first time step”

What do you mean by “first time step”? Do you mean the closest time step?

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.

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

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?

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”.

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

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.

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).

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.

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".

Figure 14: Legend is missing

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.

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).

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?

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?

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

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.