Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2019-181-RC2, 2019

Hydrology and Earth System Sciences
Discussions

Interactive comment on "A crash-testing framework for predictive uncertainty assessment when forecasting high flows in an extrapolation context" *by* Lionel Berthet et al.

Kolbjorn Engeland (Referee)

koe@nve.no

Received and published: 21 July 2019

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 transformations, 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 I list 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 underestimate 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?

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? 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.

The discussion section needs a better organization. Results presented in section 4.1 could 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.

Section 5 could be also a part of the discussion.

The number of figures could be reduced. Figure 2 – right panel is not necessary. Figure 4a and 4b could be merged. Is it possible the merge Figure 5a and b? Could results in Figure 15 be included in Figure 10? Figure 11 is not necessary. Figure 12, and 13 could 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.

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).

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?

Figure 14: Legend is missing

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.

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).”

Question or the paragraph above: which statistical model needs to be specified? Is it

for the meteorological input or for the simulated discharge?

Page 4 lines 7-8: These approaches are not exclusive of each other. Even when future precipitation is the main source of uncertainty, postprocessing 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?

Page 5 Section 2.1.1 : Maye 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.

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.

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.

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

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?

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

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."

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 the 0.1 quantile and the smallest under-estimation of the 0.9 quantile."

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."

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

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