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

The revised manuscript is much improved over the original submission. In particle, the authors have addressed my main concerns by providing more details of EHUP, better organising the discussion and (slightly) reducing the number of figures. There are still some minor issues which I feel should be addressed, so I recommend minor revisions prior to publication.

Comments

Abstract line 15: "… the Box-Cox transformation with a parameter between 0.1 and 0.3 can be a reasonable choice for flood forecasting".

This is not a key finding of the paper – the impact of using values of lambda between 0.1 and 0.3 is not shown or discussed or even referred to in the main paper, but is hidden as figure 7 in the supplementary material. Therefore, this particular finding should not appear in the abstract.

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

This an interesting question, but it does not seem a relevant title for this section since the question is not discussed.

Page 3, Lines 15-19: You mention that Renard et al (2010) models each source of uncertainty, but they do this in a hydrological prediction context (not hydrological forecasting). They do not consider uncertainty in meteorological forcing. This should be mentioned.

Page 3, Line 20: "In particular, the ensemble approaches intend to account for meteorological forecast uncertainty."

The words "In particular, the ensemble …" do not work here. I suggest replacing with "Alternatively, the ensemble …". This will nicely contrast it to the work of Renard.

Section 1.3: In this study you do not consider uncertainty in meteorological forecasts. This should be stated explicitly and justified when defining the scope of this paper.

Page 8, Lines 19-20. "Here, this highest flow group contains the top 5% pairs ranked by forecasted values."

It is not clear if this is the top 5% of all data, or top 5% of training data.

Page 8, Lines 23-24: "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."

What is the purpose of the data transformation in the EHUP? Please explain this.

Page 9, Line 13: "The Box-Cox transformation (Box and Cox, 1964) is a classic one-parameter transformation"

This statement is not really correct - the transformation you show has two parameters – lambda and a.

Figure 2 caption: "the log transformation (thick green straight line)."

This green line is not straight – it is curved.

Page 10, Lines 1-10: I still do not see the purpose of describing and providing an equation for the NQT since the NQT is not used in this study. You can explain that it is commonly used and provide reasons for why you don't use it, but no need to provide details about it.

Section 2.2.3.: You describe how G1, G2, G3 and D1, D2inf, D2sup and D3 are selected. But it is not clear how these relate to the top 5% of forecast data used for training the EHUP (Page 8, lines 9-11).

Are the data subsets based on all data, or only the top 5% of data? Or are the thresholds somehow chosen so that 5% of data is used?

Page 13, Line 14: "estimating 99 percentiles"

What do you mean by 99 percentiles? Do you mean the 99^{th} percentile (top 1%)? Or do you mean estimating percentiles between 0 and 100? If the latter, you can probably remove "99" from this sentence to make it clearer.

Section 2.3.1: What are good/bad values for each metric? You describe this for alpha-index, but not for other metrics (e.g. higher values for CRPSS are better, with 1 being perfect and 0 being same as climatology).

Figure 5: I still find this figure a bit confusing – I guess it is complicated figure since you are dealing with different periods for training (calibrating) the EHUP and calibrating the transformation parameters and evaluating the model. Here are some comments which may (or may not) help to make things clearer:

- In panel (a), it looks like the data transformation is calibrated after the EHUP is trained, which as far as I can tell is incorrect. For each set of transformation parameters, the EHUP is trained over the training period. Then based on performance over the calibration period, the set of calibrated parameters is chosen. So panel (a) is really about calibrating the transformation parameters.
- In panel (b), you are re-training the EHUP with the calibrated set of transformation parameters and additional data, and then evaluating performance using another set of data. Perhaps this can somehow be made clearer in the figure/caption.

Figure 5 caption: "The straight lines represent their use to assess …"

Do you mean the vertical lines?

Figure 5 caption: What are the grey dots?

Page 18, Lines 9-10. "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."

This should be elaborated on. You provide details about this in the figure caption, but they should be provided in the text.

Section 3.2.2 heading: "Overall performance"

This section describes "overall performance" using the CRPSS metric. But it also describes accuracy and sharpness. The heading is misleading since it only mentions overall performance. Perhaps change to "Other performance metrics"

Page 20, Line 11: Reference to Supplementary Material**.** Please provide reference to specific figure numbers in supplementary material.

Same comment applies for other references to Supplementary Material.

Section 4.2: The results presented in this section are somewhat interesting, but I do not see how they address a specific aim of the paper. As such, they do not seem to add value to the paper and in my opinion are a distraction from the key findings of the paper.