Revisions for the manuscript "evalhyd v0.1.1: a polyglot tool for the evaluation of deterministicand probabilistic streamflow prediction" by Hallouin et al.

General comments

This article presents a hydrological verification tool which is coded in several languages to accommodate different users used to different coding habits. The tool includes some data pre-processing capabilities, as well as inference by bootstrapping and the possibility of applying conditional verification (e.g. tailoring different streamflow regimes). The article reads very well and is logically structured; the language, figures and text are very clear; the results shown corroborate the conclusions.

I really enjoyed reading this article! and I have no hesitation in recommending it for publication. Here below I outline some minor suggestions, for the author's considerations, after which the manuscript can be published.

Specific comments

1. Line 105, the dimensions for a deterministic evaluation *{series,time}* seems to me inconsistent with the dimensions for the probabilistic evaluation *{sites, leadtimes, ensemble members, time}*, introduced at line 99. I would add "*sites, leadtimes*" also for the deterministic evaluation to have *{sites, leadtimes, series, time}*. (essentially you have replaced the "*series*" of multiple simulations to "*ensemble members*")
2. I have few questions about the bootstrapping: at lines 166-167 I understand that each year is a block, but I do not understand what are the sub-periods: can you please explain. Do you perform a resampling with replacement? (allowing a year to be selected multiple times in the same bootstrap sample). At lines 183-184 you mention that the user can evaluate a custom-made skill score with the reference of his/her choice (and that evalhyd evaluates the two scores separately, prior evaluating the skill score): do you use a pair-bootstrapping for the confidence intervals on this skill score, right? (aka the bootstrap re-samples are the same for the prediction and the reference).
3. In the example showcased in section 6 you compare the prediction against persistence. Can you please specify what is the persistence forecast: is it a fix persistence (e.g. for each validity date you consider as prediction the streamflow of n days before, with n fix), or does the persistence align with the prediction lead-time, so that a 10 day forecast is compared to a 10 day persistence, and a 2 day forecast is compared against a 2 day persistence? (essentially the prediction is compared to its initial state). Thank you for specifying this in the article.
4. Figure 4 and A3: the CRPSS against climatology consistently decreases with lead-time (as expected), for almost all stations, whereas the CRPSS against persistence does not: why? Is there a sort of re-emergence of the signal which renders the persistence more skillful for some time-lags?
5. Table 2: please define the Mean Absolute Relative Error (or provide a reference).

6. Why for the deterministic prediction the contingency table entries are provided (Table 2 last line), whereas for the probabilistic prediction (Table 3) you compute directly POD, POFD, FAR, CSI? Would not be more symmetric to provide also for the probabilistic metrics the contingency table entries?
7. **Conditional masking: in light of the comments in the main discussion item on conditional verification, please consider adding some text in the article about the effects of bilateral versus unilateral conditions when performing conditional verification.**
8. Figure 7 shows regional scores for the river basins: The Loire basin seems performing very well, however in Fig.3 (and A2) if was shown that for two of the upstream stations the skill was quite poor. Maybe we should not aggregate regionally? I suggest to remove this option and section 6.2.3.

Technical corrections

1. Especially at the beginning of the article, it is not very clear what is meant with "post-processing of the computed metrics" (it becomes clear only after reading Section 6). Since the term "post-processing" is usually used for statistically correct a forecast, I suggest using different phrasing.
    a. Line 3 can be rephrased as " … , but it can also involve the data quality control and pre-processing, as well as performing further analysis (e.g. tailored stratification, inference) on the computed metrics".
    b. Line 29 can be rephrased as " … , the metrics can be subject also to sensitivity analysis and uncertainty estimation"
2. In section 3.1 (e.g. lines 72, 80, 83) you talk about "satellite bindings": can you please avoid using "satellite", and maybe just refer to these as "bindings" or (even better) "extensions" (as they are named `xtensor-python, xtensor-r,` etc.)
3. Line 111-112: the "the quadratic error between the observations and their arithmetic mean" is the observation variance, right? I suggest mentioning this, e.g. in a parenthesis following the text.
4. Lines 123-128 need rephrasing, I'll try rephrase some text (as I understood the issue), e.g. line 124: "In addition, when predictions for several lead-times are considered, the predictions with validity time beyond the observed timeseries need to be flagged as *not a number*." At line 126 I suggest not using "invalid dates" but rather write " … and dates where the observation has no matched prediction must be identified as *not a number*". I also suggest referring to *initialization + lead time = validity time* for the prediction to be matched with the observation at the validity time.
5. The caption of Fig.2 need rephrasing too, in view of the previous comment. Moreover, it is the last row which shows the observation (and not the first), and the prediction is on the preceding (and not following) rows. What are the numbers within the rounded rectangles?

6. Lines 180 and 182: the reference for the NSE is indeed the mean of the observations (aka the sample climatology). The reference for the BSS is (correctly) again the sample climatology. However, from the way these sentences are phrased, these two references seem different references. You might want to rephrase your text to fix this, e.g. line 180 could be " … where the references is taken as the mean of the observations (aka the sample climatology)"
7. Caption of figure 6: please specify that the condition is applied only for the forecast, e.g. you can write " (a) for predicted low-flow conditions, … ".
8. Line 268: use the present tense "is presented" (and not "was presented").