Dear Editor, Dear Referees,

We would like to thank you for considering our manuscript, for the thorough review, and for the constructive feedback you provided. Please find below a detailed point-by-point answer to the referees' comments.

The authors.

Nomenclature:

RXCY – Referee number X Comment number Y
AR – Authors' Reply
O-LX – Original manuscript Line X
R-LX – Revised manuscript Line X

Referee #1 (Barbara Casati)

General comments

R1C1: I really enjoyed reading this article, and *evalhyd* seems a very nice verification too: felicitations! I have uploaded a file with some minor suggestions, while here with this following comment I wish to bring the attention of the scientific community to one aspect of verification, for the general online discussion.

I was particularly triggered about the option in *evalhyd* of performing conditional verification. I will share here some of our own experience (at the Canadian Met Service) with conditional verification, which maybe can inspire further developments in the tool and, more in general, awareness in the interpretation of the results.

Conditioning on the verification sample can have strong impacts on the verification results (e.g. it can flip the sign of a bias), and hence allows in-depth analysis and understanding of the prediction performance, since the conditioning is usually related to physically-driven phenomena. In a sense, conditional verification is the first step towards process-based diagnostics.

In verification exercises which include several variables (e.g. pressure, temperature, clouds, etc) applying a condition to a variable while verifying a different variable is the common practice (as an example, verification of surface temperature in cloudy versus clear-sky conditions inform of the model performance in reproducing the radiation budget). The condition, however, should be applied to both observed and forecast values (e.g. forecast AND observation being cloudy): I will refer to this double condition as *bilateral*. On the other hand, when a *unilateral* condition is applied, to only the observed or forecast variable (e.g. cloudy conditions only for the forecast) this can synthetically introduce a bias in the verification results: in the cloud/temperature example above, stratifying for cloudy conditions only in the forecast leads to a synthetic warm bias for the surface temperature, because in the sample there are bound to be both cloudy

and clear sky observations, and when the observations have clear sky the surface temperature is expected to be colder. In other words, the bilateral condition will sample all the "hits" for cloudy sky, whereas the unilateral condition will sample the "hits" and "false alarms" for cloudy sky. From our experience, we advise bilateral over unilateral conditioning. (Of course one can also do unilateral conditioning, but need to be aware of the introduced biases in the interpretation of the verification results).

Applying the unilateral condition to the same variable which is verified might also lead to synthetic biases. As an example, if you stratify your sample for the strong *predicted* stream flows, you are bound to include in the sample several strong observed stream flows (the "hits"), but also some average or weak observed stream-flows (because the prediction might have some "false alarms"). Then you tend to "artificially" diagnose over-prediction for the strong stream flow (and vice-versa for the low stream flow, conditioning only on the prediction you are bound to find under-estimation, because in your sample you'll have some observed events which are medium or strong).

(...)

I would be grateful if you could add in the article some discussion about unilateral versus bilateral conditions.

AR: Thank you for sharing your experience on conditional verification and for the very detailed explanations on the potential biases that unilateral conditioning can lead to. We have added some references in the introductive paragraph of the masking functionality section (see R-L137-140) to give some context on the motivation behind conditional verification. In addition, we have added a paragraph in the limitations of the tool mentioning the risk of introducing synthetic bias when performing unilateral masking and/or applying the condition on the variable being evaluated (see R-L290-301), and we have added a cross-reference to the limitations section in the masking functionality section for the reader not to miss them. We hope that these additions are faithful to your explanations.

R1C2: I was amazed (also a bit puzzled) to see that in your Figure 6 you have opposite results than I expected (underprediction for high predicted stream flow, more overprediction for low predicted stream flows; the under-dispersion for the average predicted stream flow is instead expected). For me it would be interesting to understand why, is it due to the characteristics of streamflow prediction (where the timing is always predicted well, and hence false alarms and

misses are very rare)? what is the behaviour in the other stations? What would you obtain with the bilateral condition?

AR: This figure seems relatively coherent with what is expected given the difficulties in capturing the hydrological variability: models tend to overestimate low flows and to underestimate high flows. In order to provide in-depth explanations for these results, it would be necessary to also consider rainfall predictions to be able to distinguish the quality of the model from the quality of the predictions. This goes beyond the scope of this technical article that only intends to demonstrate what can be produced with the tool, and not to analyse the reforecasts used as an example data set that is well-known and openly accessible.

Specific comments

R1C3: 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")

AR: As mentioned in section 3.3, the four-dimensional character of the inputs for the probabilistic evaluation is motivated by the existence of multi-variate probabilistic metrics (e.g. the energy score proposed by Gneiting et al. (2008)). This is not commonly the case for deterministic evaluation, as there is by definition only one forecast member, and there is no multi-site metric in common use to the best of our knowledge. This is why we decided to keep the dimensionality of the inputs as low as possible. Indeed, deterministic metrics are often used in optimisation problems, where one-dimensional inputs would be sufficient (to compare one simulation series against one observation series), so it appeared to us that it would be heavy for such case to include so many extra dimensions, potentially confusing users only interested in determistic evaluation and not familiar with ensemble forecasting concepts. The "series" dimension can perfectly be used to provide deterministic forecasts for several lead times. But, as mentioned above, the use of the term "leadtimes" for the deterministic evaluation would be reductive to a sole forecast context. So we believe that retaining the dimensionalities as they currently are is the best compromise. We have added forecasting mentions in this section for completeness and clarity (see R-L110-113)

R1C4: 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).

AR: The sub-periods are made of randomly drawn year blocks. In the revised manuscript, we have reformulated our explanations in the text (see R-L178-181) and added a figure to illustrate the bootstrapping functionality (see Figure 4 in the revised manuscript). Yes, the tool does perform a sampling with replacement. We followed the recommendations made by Clark et al. (2021) to use a non-overlapping block boostrapping, i.e. to draw complete years with replacement. We have added this information to the manuscript (see R-L179-180).

R1C5: 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).

AR: Regarding custom-made skill scores, you are absolutely right in wondering whether the sampling is consistent between the computation of both scores before combining them, as we believe it should be. Thank you for raising this. The tool does allow to keep the same sampling by setting the seed of the pseudo-number generation used for the random sampling with replacement. So, while this cannot be enforced by the tool, this is possible (and essential) to set the same seed when computing both metrics. We have added this information and this recommendation to the manuscript (see R-L203-206).

R1C6: 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.

AR: The persistence forecast benchmark corresponds to the latter, i.e. the prediction is compared to the same initial conditions for all lead times. We have specified this in the article by directly quoting Harrigan et al. (2023) and, for completeness, we have also added details about the climatology benchmark (see R-L214-218).

R1C7: 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?

AR: This behaviour is discussed in detailed in the peer-review discussion of Harrigan et al. (2023), it is accessible at https://hess.copernicus.org/preprints/hess-2020-532/hess-2020-532-AC1-supplement.pdf (last access: 21 Jan 2024, see discussion for the second specific comment). In short, this is due to the fact that, while both the reforecast and the persistence benchmark deteriote with increasing lead times, the decline in the persistence accuracy happens at a much faster rate than the accuracy of the reforecast. Again, we believe that these aspects are important but they go beyond the scope of a technical paper such as ours.

R1C8: Table 2: please define the Mean Absolute Relative Error (or provide a reference).

AR: This corresponds to the Mean Absolute Error (MAE) divided by the observed mean, we have added this information as a table footnote (see Table 2 in the revised manuscipt)..

R1C9: 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?

AR: Yes, we agree with that. We have added the contingency table as a metric (CONT_TBL) in the probabilistic entry point of evalhyd (available in v0.1.2). Table 3 in the revised manuscript now features CONT_TBL as an available probabilistic metric. Thank you for the suggestion.

R1C10: 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.

AR: We have addressed this comment together with the first comment, please refer to our reply for **R1C1**.

R1C11: 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.

AR: We agree that multi-variate scores such as the energy score can hide poorly predicted stations, however, this is the drawback with any aggregation approach, as is the case when interpreting results for large sample hydrology for instance. While such metric should be employed with care, we do believe that it can have some relevance in some applications. In particular, unlike a straightforward mean of CRPS values across stations, the Energy Score employs a weighted Euclidian distance between the different stations considered. We verified our results here and decided to maintain this section for these reasons.

Technical corrections

R1C12: 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.

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".

Line 29 can be rephrased as " … , the metrics can be subject also to sensitivity analysis and uncertainty estimation"

AR: Thank you for highlighting this potential confusion due to our choice of words. Throughout the manuscript we have replaced "pre-processing" by "prelimary processing" and "post-processing" by "subsequent processing", while providing early on in the introduction examples of each type of processing (see R-L22-24).

R1C13: 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.)

AR: Given the geoscientific scope of the article, we understand that the use of this particular word figuratively may have been unwise. We followed your piece of advice and dropped satellite in favour of simply "bindings" (see R-L73-90). Since "extension" implies the addition of functionality, which is not the case here, we preferred not to use this term.

R1C14: 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.

AR: That is right. We have simplified by replacing it directly by the "observed variance" (see R-L118), thank you for the suggestion.

R1C15: 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.

AR: This is indeed a part we found tricky to formulate. Thank you for your suggestions. The term "valid date" was inspired by the CF conventions (https://cfconventions.org/Data/cf-conventions/cf-conventions-1.11/cf-conventions.html#scalar-coordinate-variables, last access: 16-12-2023): "Multiple forecasts from a single analysis (…) The analysis time is identified by the standard name "forecast_reference_time" while the valid time of the forecast is identified by the standard name 'time'." In fact, "valid_time" is the field used by ECMWF in their forecast. However, for clarity, we propose the following rephrasing: "Therefore, when several lead times are considered at once, a temporal shift of the predictions must be applied, and observed dates for which a forecast is not made (i.e. where date ≠ forecast issue date + lead time) must be identified as *not a number*." (see R-L131-133).

R1C16: 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?

AR: The caption was rephrased accordingly, as per our reply to **R1C15**: "before and/or after the prediction series to align the prediction validity dates (i.e. issue date + lead time) with the observation dates when several lead times are considered at once". We have also fixed our mistake with the relative locations of the predictions and the observations (the former is indeed below the latter). Thank you for spotting this.

R1C17: 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)"

AR: Yes, this is worth using the same terminology between those two references indeed. As suggested, we have added in parentheses for the NSE that its reference is the sample climatology (see R-L196).

R1C18: 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, ... ".

AR: Yes, as we performed unilateral conditioning, we have prefixed with "predicted" for sub-labels (a), (b), and (c) of Figure 6 in the revised manuscript.

R1C19: Line 268: use the present tense "is presented" (and not "was presented")

AR: This is done (see R-L303).

Referee #2 (Anonymous)

General comments

R2C1: The paper introduces evalhyd, an interesting software tool designed for the evaluation of streamflow predictions. The tool's commitment to standardization and open-source accessibility is commendable, providing a valuable contribution to enhancing reproducibility in hydrological studies. Notably, the well-thought-out design principles, which incorporate a compiled C++ core and thin bindings for multiple languages, contribute to the tool's efficiency and usability.

However, despite the paper positioning evalhyd as a contribution to hydroinformatics, the manuscript's focus on the technical aspects of model development limits its scientific impact. The paper could benefit from a more explicit emphasis on the broader scientific implications and advancements in hydrologic science that the tool facilitates.

(...)

Overall more critical analysis is needed on how evalhyd improves on the current state of the art in hydrologic evaluation tools. The paper currently lacks motivation and innovation.

AR: We thank the reviewer for the comments and suggestions. The tool developed contributes to advancing best practices used in hydrological evaluation by making available a tool that features advanced methods seldom used in the hydrological community, and by simplifying their widespread adoption by providing an efficient and easy-to-use tool that is usable in the main programming languages used by the community. We believe that, while this is not the end of the road, this represents a far from negligible step towards achieving reproducible science in hydrology and, as such, is a valuable contribution to hydrological science. However, we agree with the reviewer that this was not explained well enough in the original manuscript, therefore we have extensively reworked the introduction to emphasise these motivations and innovations around our tool (see R-L25-32, R-L50-59).

Specific comments

R2C2: The introduction would benefit from more clearly articulating what new capabilities evalhyd provides compared to existing hydrologic evaluation packages. As it stands, the motivation around standardization across languages is a bit weak.

AR: We modified the introduction accordingly. Please see our reply to **R2C1**.

R2C3: In the key functionalities section, the masking and bootstrapping methods need more detailed explanation. Pseudocode or formulas would help make these clearer.

AR: Thank you for raising this issue and for the suggestion. We considered using pseudo code and formulas, but we decided that figures displaying trivial examples of both functionalities would be more understandable. This is why we have produced two additional figures were produced that aim to complement the explanations for these two key functionalities (see new Figures 3 and 4 in the revised manuscript). For the sake of coherence, we took care of using similar symbology with the existing figure on the handling of missing data.

R2C4: For the evaluation metrics, links or references to the original sources for each metric should be provided. More justification for the specific metrics included would also help show the comprehensiveness.

AR: Tables 2 and 3 in the original manuscript provide the original sources, where possible. However, given the general nature (i.e. standard statistical metrics not always specific to hydrology) of some of them, references are not always possible to find.

R2C5: The case study, more novel demonstrations of the tool would strengthen this section.

AR: We do believe that the existing demonstrations of the stratification (i.e. masking) functionality or the bootstrapping functionality represent novel and advanced methods for the evaluation of hydrological predictions. The focus of our manuscript is the development of the tool and its demonstration (using data from an independent already published paper) and, given its numerous fonctionalities, additional demonstrations would take a lot of space in the paper and change its focus.

R2C6: The conclusions would be improved by specifically emphasizing the limitations around extensibility, visualizations, and support for continuous distributions. Comparisons to other existing packages may help contextualize the pros/cons.

AR: Thank you for the suggestion, this is indeed worth recapitulating the limitations as well. We have added a paragraph summarising the main limitations in the conclusions (see R-L311-312). The comparison to existing packages is already made in the paragraph before last in the introduction (see O-L32-48, or R-L33-43).

References

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., et al. (2021). The abuse of popular performance metrics in hydrologic modeling. Water Resources Research, 57, e2020WR029001. https://doi.org/10.1029/2020WR029001.

Gneiting, T., Stanberry, L., Grimit, E., Held, L., and Johnson, N. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds, TEST, 17, 211–235, https://doi.org/10.1007/s11749-008-0114-x.

Harrigan, S., Zsoter, E., Cloke, H., Salamon, P., and Prudhomme, C. (2023). Daily ensemble river discharge reforecasts and real-time forecasts from the operational Global Flood Awareness System, Hydrology and Earth System Sciences, 27, 1–19, https://doi.org/10.5194/hess-27-1-2023.