Comment on hess-2022-141
Anonymous Referee #3

Referee comment on "Low flow estimation beyond the mean – expectile loss and extreme gradient boosting for spatio-temporal low flow prediction in Austria" by Johannes Laimighofer et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2022-141-RC3, 2022

GENERAL COMMENT

The manuscript by Laimighofer et al. presents a new and interesting approach for monthly low-flow estimation, based on the use of a statistical learning model (i.e. the extreme gradient tree boosting model XGBoost). It is presented an analysis on a rather large dataset of 260 gauging stations in Austria. One of the major element of novelty is the use of the expectile loss function as a fitting criterion; and, in particular, different expectiles (from 0.01 to 0.5) have been applied and tested. In my opinion, the proposed method contributes relevantly to the research field focused on the prediction of low flow in ungauged catchments. The topic fits the purposes of Hydrology and Earth System Sciences and it is currently in fashion among many research groups owing to the potential relevance for climate change that could profoundly modify low flow timing and magnitude in the next future.

The core theme of the manuscript has merit. Overall, this paper is logically organized and English language and style are fine. The methodology and procedures adopted are appropriately described. The authors do an admirable job of making plain a topic that can be incredibly dense and difficult to understand. Results are clearly discussed and the conclusions are adequately supported by the analysis presented.

Nevertheless, I have some suggestions, general comments and specific remarks, which I believe should be addressed prior to publication. Given their nature, I believe that a minor revision is needed for addressing all of my concerns, reported below. Although my review may seem rather critical, I think this is a very good paper. I hope that my suggestions could serve to ameliorate the paper. Best regards.
SPECIFIC COMMENTS

Sect. 1. Introduction

L.40 p.2 – “...need of less data...”

I respectfully disagree about this, since data-driven models always require a large amount of data (longer periods of observations for the response variable compared to other approaches).

Sect. 2.1.1 Hydrological Data

L.76 p.3 – “The dataset was already used in a wide range of studies (e.g. Laaha and Blöschl, 2006, 2007; Laaha et al., 2014; Laimighofer et al., 2022). Some stations included in previous studies had to be discarded as the gauging stations were removed, relocated or the data included too many missing values.”

These two sentences are rather unclear to me. How is this information relevant to this study? Why refer to previous studies here? In my opinion, what is important here is to define the dataset and the preliminary analyses considered for the present study.

Sect. 2.1.2 Predictor variables

This is actually my major concern: the use of symbols and acronyms could be greatly improved throughout the manuscript, and, in truth, I found Sect. 2.1.2 rather confused. I would suggest carefully reviewing the manuscript, renaming some symbols, defining ALL the adopted symbols at their first appearance in the manuscript, providing description of some indexes if needed, and homogenizing all, avoiding the use of different symbols for
the same thing.

For example:

- At L.103-104 p.5. \( P_{\text{in}} \) and \( P_{\text{um}} \) should be defined; the acronym MCBW and the various subscripts should be explained (what does the “M” mean? what do “in” and “um” mean?).
- In the caption of table 1, the subscripts “win” and “sum” are explained (they should be defined also in the text at their first appearance), but annual partitioning into two seasons has never been discussed. “Summer” is not properly the summer, but it seems to be a six months period whose start and ending times are unknown (they should be specified). The same for the “winter”. Why not use the terms “dry” and “wet”? In my opinion they are better suited to the bi-seasonal division of the year.
- In table 1, the description of some variable should be expanded. For instance \( S_{\text{SL}} \), \( S_{\text{MO}} \), and \( S_{\text{ST}} \): these three classes of slope should be defined. Have you considered some kind of slope threshold value or something else to classify the slope? Please, specify.
- At L.109 p.6, the symbol for the monthly climatic water balance becomes CWB. Is this different from MCWB? If not, you should homogenize the use of symbols.

L.107-111 p.5-6. Here a preliminary assessment is mentioned. Please indicate how do you perform this preliminary assessment and the metric adopted to evaluate the performances.

L.119 p.6. The standardized drought index should be defined. How is it computed?

L.121 p.6. “Fig.1” should be Fig.2

Sect. 2.2.1. Extreme gradient tree boosting

How were the range of variation of the hyperparameters (e.g. maximum depth, etc...) reported between L.146 and L.149, p.8), selected? Do they come from literature? Are they default or typical ranges? What else? Please, specify.

L.151. (p.8). Please, specify what you exactly mean with 10-fold CV. What does the symbol CV mean? Cross-Validation? Please, add explanation to the manuscript.
2.2.2. Loss function.

L.167 p.8. Double “of”

Sect. 2.2.3. Variable selection

L.185 p.9. Are “10 CV” and “10-fold CV” the same thing? If so, please homogenize the symbols, otherwise, explain the differences.

L.85-186 p.9. How was the threshold of 1.05 selected? Please, specify

Sect. 2.2.4. Model evaluation

L.190 p.10. $L_{\text{MDAE}}$ was already explained in Sect. 2.2.2. Why introducing another symbol (MDAE) for the same thing? The same for $L_{\text{MAE}}$. Consistently with the symbology adopted in the present study, “RMSE” in eq.5 should be $L_{\text{RMSE}}$, and the coefficient of determination in eq. 6 should be $L_{R^2}$

Sect. 3.1. Global model performance

Here the authors focus almost entirely in the description of the results reported in Tab. 3. I would suggest trying to expand the discussion, addressing some observed behaviors.
Sect. 3.2. Station-by-station performance

L267-269 p.13. “However, the 0.2 and 0.3 expectile still show better performance than the mean absolute and median absolute loss in terms of the $R^2_{\text{med}}$ and also a low portion of stations with weak performance (only 36% (0.2) and 32% (0.3) stations have a $R^2$ below 0.5).”

Please, quantity the “better performance” reporting the associated values of $R^2_{\text{med}}$ or recalling Figure 5 or Table 4 in the sentence. “Portion” at L.268 should be “fraction” or “percentage”. Avoid the use of nested round brackets (..(..)..).

Looking at figure 5, Expectile 0.01, 0.025 and 0.05 should be discarded since they provided very few (or none for Expectile 0.01) stations with acceptable values for $R^2$. Also the other metrics in Tab 3, at global level confirms their scarce suitability. This should be better highlighted.

Sect. 3.2.1. Error decomposition

I found this analysis very interesting, and in my opinion the discussion should be expanded a little bit trying to interpret and address the results. For example, looking at Figure 6, it seems that the main difference among the various expectile models is the produced bias, while the prediction seasonal and annual errors provide about the same relative contribution to the total error. Why? Was it expected? I would suggest to discuss the main implications in considering models with different expectiles (i.e. Expectile 0.5 instead of Expectile 0.01).

Sect. 3.3. Prediction of extremes
L.298-300. p15. "For example if we assess the accuracy of our models for cases below the 1 % quantile, the 0.01 expectile is yielding a $R^2$ of 0.42, where larger expectiles (0.1 - 0.5) show inefficient models with a $R^2$ below 0."

In my opinion also a value of $R^2$ of 0.42 could denote inefficient models, and, indeed, values around 0 (both negative or positive) are all representative of inefficient models and should not be compared with each other. For example, I am not sure that a model with $R^2$ of 0.1 is better than a model providing a negative value for $R^2$, while I rather confident that both don't work.

L.318 p.17- L 330 p.18. Last part of Sect.3.3

I would suggest to provide a possible explanation that may justify the observed increasing trend for the hit score and decreasing trend for the precision index with decreasing expectile (from 0.5 to 0.01) and equal quantile. The authors should probably consider the definitions of the two metrics (see eqs. 12 and 13), their meaning and implications in terms of models prediction ability. Looking for an optimal trade-off solution, which model should be preferred? Which metric carries the highest weight in describing models prediction ability considering the different target that the model may have?

In the last sentence the authors state "...the user needs to find some optimum“(L.330), but in my opinion authors should provide here some practical indications to do so. This important aspect is roughly mentioned only in the conclusion section.

Sect.3.4 Variable selection

L.340 p.19. Please, explain what you exactly mean with "...variable importance of less than 2 %...". Which metric is considered in this analysis to assess the % gain?

Sect.4.2 Performance compared to literature
First paragraph (from L.379 to L.393. P.22). Here, it seems that the authors compare quantitatively different metrics (i.e. coefficient of determination and Nash and Sutcliffe Efficiency). I think that the comparison can be done quantitatively only using the same metric, otherwise it can be done only qualitatively, for example considering some performance rating able to classify the different performances (e.g. “weak”, “good”, excellent”, etc.), which usually are specific for the variable under consideration, the adopted metric and the time and space scale of analysis.

L.388 P.22. “Sicilia” should be “Sicily”

Sect.5 Conclusion.

L.418-420 P.23. Given the high number of predictor variables considered, could overfitting partially explain the overall performance drop for extreme low flow and for decreasing expectiles (e.g. 0.01 or 0.025)? Perhaps, the authors might add something about this “potential” problem and how they tried to prevent overfitting.