

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1 https://doi.org/10.5194/hess-2021-65-RC1, 2021 © Author(s) 2021. This work is distributed under the Creative Commons Attribution 4.0 License.

## Comment on hess-2021-65

Anonymous Referee #1

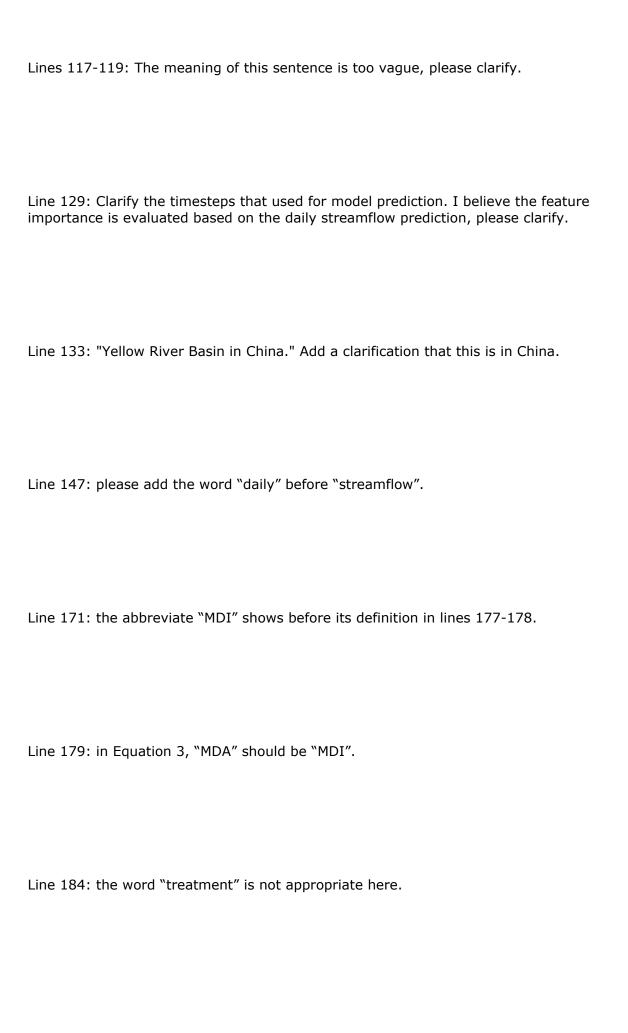
Referee comment on "Development of a Wilks feature importance method with improved variable rankings for supporting hydrological inference and modelling" by Kailong Li et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-65-RC1, 2021

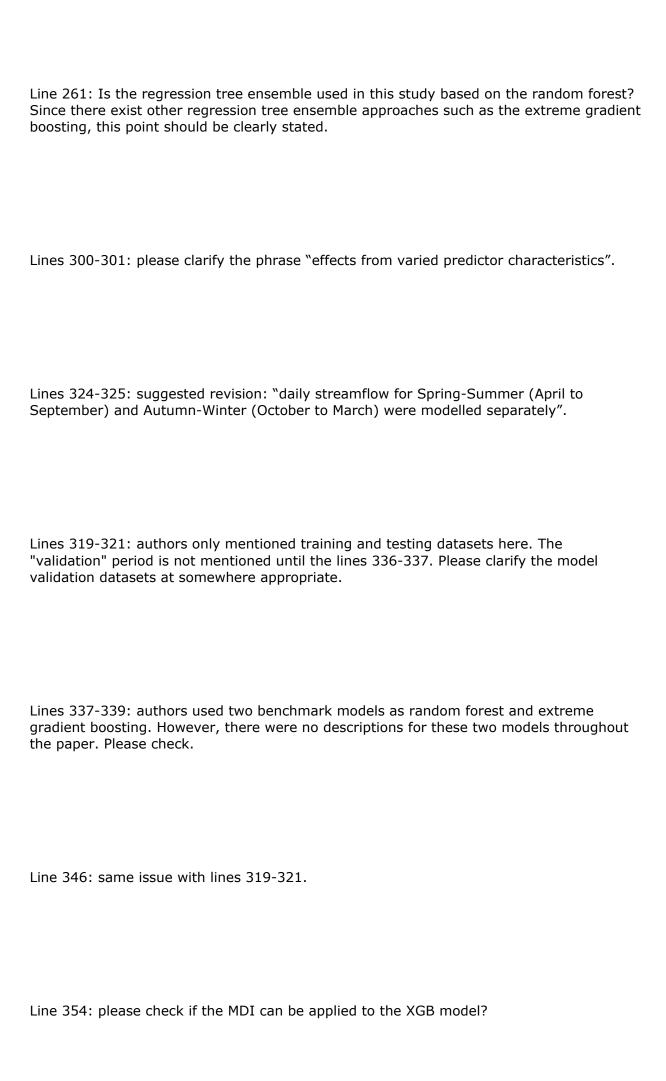
The authors developed an interesting approach of quantifying the relative importance of predictors for decision trees. Overall, the manuscript is well-organized, and the method and analysis appear to be well developed. To my knowledge, the interpretable machine learning has been receiving increasing attentions from hydrological community, and the proposed method would contribute to such a growing body of research. I have some concerns as indicated in the comments to the authors.

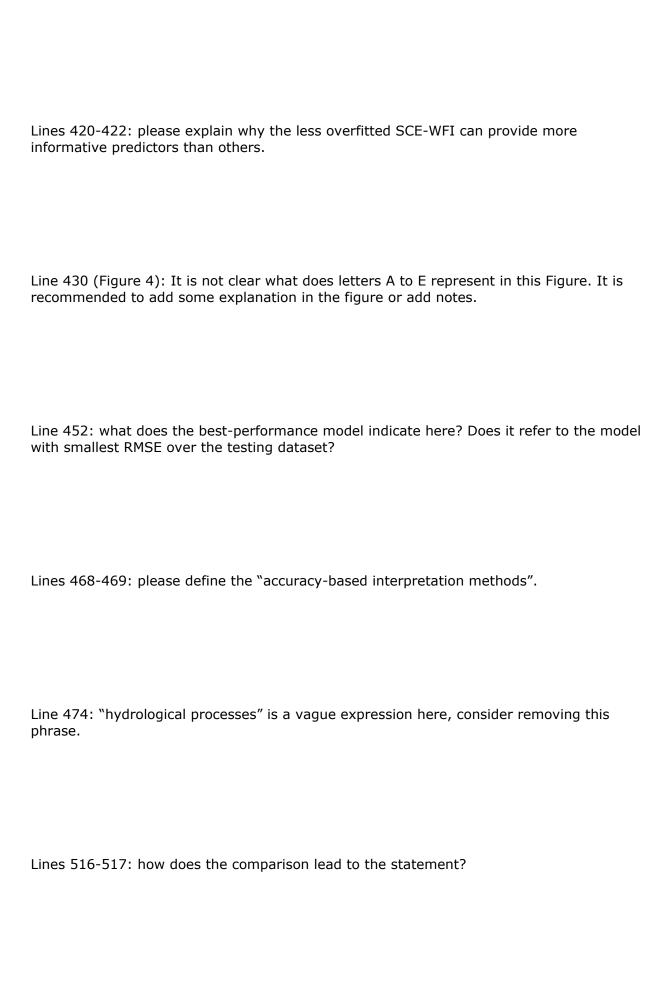
General comments:

The new method proposed in this study seems to quantify the importance of predictors from the training datasets and then validate the importance over the testing datasets through two evaluation metrics. The authors argue that the proposed method can identify key predictors that may be overlooked by other interpretation methods. By testing this hypothesis, those "overlooked" predictors are reconsidered into models, and this leads to improved accuracy on the testing dataset. The question is that the authors only tested that conventional interpretation methods may overlook some predictors but did not test whether there exist same isues regarding the proposed method. For instance, in Figure 7, the winter irrigation period of the 3<sup>rd</sup> irrigated watersheds, I7 was considered as important predictors by other models but not by the proposed method. So how will the model perform without this predictor? Similar concerns also exist for the other irrigated watersheds and irrigation seasons. Please list more evidences that the proposed methods can select more informative predictors.

Make it clear that what datasets are used for training each of the interpretation methods. I have seen many literatures using testing dataset to evaluate the feature importance, which I think is not appropriate. In this study, the authors seem used different datasets (training and validation) to calculate the importance scores under different interpretation methods. This should be clarified. In addition, the validation datasets used in this study were not described clearly.
There are many small errors throughout the manuscript and need to be corrected. For example, authors sometimes use "irrigated watersheds" and sometimes use "drainage basins", similar phrases should be unified.
Specific line comments:  Lines 20-21: make it clear that the predictors identified by WFI using the training dataset achieved the highest predictive accuracy on the testing dataset.
Lines 24-25: the related findings should be more specific.
Line 87: the regression tree ensembles are not necessarily composed of hundreds of interpretable models (i.e., decision trees). It is suggested to add the word "usually" before "composed" for clarity.







Lines 733-735: clarify that the simulation period is Spring irrigation in Table 3.