Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/hess-2021-65-RC1, 2021

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 3rd 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 117-119: The meaning of this sentence is too vague, please clarify.

Line 129: Clarify the timesteps that used for model prediction. I believe the feature importance is evaluated based on the daily streamflow prediction, please clarify.

Line 133: "Yellow River Basin in China." Add a clarification that this is in China.

Line 147: please add the word “daily” before “streamflow”.

Line 171: the abbreviate “MDI” shows before its definition in lines 177-178.

Line 179: in Equation 3, “MDA” should be “MDI”.

Line 184: the word “treatment” is not appropriate here.

Line 261: Is the regression tree ensemble used in this study based on the random forest? Since there exist other regression tree ensemble approaches such as the extreme gradient boosting, this point should be clearly stated.

Lines 300-301: please clarify the phrase “effects from varied predictor characteristics”.

Lines 324-325: suggested revision: “daily streamflow for Spring-Summer (April to September) and Autumn-Winter (October to March) were modelled separately”.

Lines 319-321: authors only mentioned training and testing datasets here. The "validation" period is not mentioned until the lines 336-337. Please clarify the model validation datasets at somewhere appropriate.

Lines 337-339: authors used two benchmark models as random forest and extreme gradient boosting. However, there were no descriptions for these two models throughout the paper. Please check.

Line 346: same issue with lines 319-321.

Line 354: please check if the MDI can be applied to the XGB model?

Lines 420-422: please explain why the less overfitted SCE-WFI can provide more informative predictors than others.

Line 430 (Figure 4): It is not clear what does letters A to E represent in this Figure. It is recommended to add some explanation in the figure or add notes.

Line 452: what does the best-performance model indicate here? Does it refer to the model with smallest RMSE over the testing dataset?

Lines 468-469: please define the “accuracy-based interpretation methods”.

Line 474: “hydrological processes” is a vague expression here, consider removing this phrase.

Lines 516-517: how does the comparison lead to the statement?

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