Reply on RC1

Kailong Li et al.

Author 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-AC1, 2021

Dear Anonymous Referee #1,

Thank you very much for your time and efforts to review our manuscript. Your suggestions are very valuable and helpful to improve our manuscript. In this short response, we will briefly address your major concerns under the General comments. A detailed point-by-point response will be provided later, along with the revised manuscript.

The first major concern is related to the robustness of the proposed feature importance method (i.e., whether it can achieve more reliable importance scores than other methods).

In the original manuscript, we compared the difference of key predictors identified by three different feature importance methods, and found the predictors identified by the proposed WFI method led to the highest predictive accuracy not only to the SCE model but also to the RF model over the testing dataset. However, we only verified the proposed method on three watersheds, and realized that the sample size is inadequate. In fact, this concern was also raised by Anonymous Referee #2. Therefore, to address the credibility of the proposed method, we will run much larger datasets, including 20 more basins with various flow characteristics. We believe that a much larger dataset will significantly improve the credibility of the proposed method.

The second major concern is about the clarity of datasets used under training, validation and testing procedures. In the revised manuscript, we will clarify all the datasets and associated purposes. In short, both MDI and WFI methods use training dataset to calculate the importance scores, while the PFI method uses the validation dataset to calculate the importance scores. The training dataset is obtained from sampling from the original training set with replacement (the sample size is as same as the original training set). This process can leave about 1/3 of the training dataset as out-of-bag (OOB) data; these OOB data will not be involved in the model training process, and can thus be used for validation.

Again, we much appreciate the reviewer's careful review for pointing out some inappropriate expressions in the manuscript. In the following revisions, we will carefully check our manuscript to address these concerns.

Best regards,

Kailong Li, on behalf of the team of authors