Reply on RC2

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-AC2, 2021

Dear Anonymous Referee #2,

We much appreciate your time and efforts in reviewing our manuscript. Your constructive and encouraging comments are very helpful for better clarifying the focus of our study. In this short response, we will briefly address your major concerns. A detailed point-by-point response will be provided later, along with the revised manuscript.

We much agree with your suggestion and will reframe the introduction section as follows: it will start with (1) providing evidence that the interpretable machine learning has been receiving increasing attention from the hydrological community, followed by (2) detailed information on existing feature importance methods, splitting rules, and potential issues, after that (3) new feature importance methods and splitting rules will be provided and state that how the existing one can be improved in summary, and finally (4) real-world hydrological applications will be introduced.

In order to prove that the method performs well compared to existing ones, 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. In addition, we will perform several comparative studies (e.g., analysis of single tree and multiple trees under two different splitting rules) to illustrate how and why the proposed method can outperform the existing method.

We will remove the equifinality principle from our introduction in order to clarify our goal of this study and improve the readability of the manuscript.

We regret the unclear definition of some concepts (e.g., equifinality, interpretability, collinearity and predictivity). In the revised manuscript, we will clarify these concepts and try to avoid confusion.

Bayesian model averaging (BMA) used in this study is a post-analysis approach to the proposed feature importance method (i.e., WFI). BMA further investigates how the importance scores obtained by WFI vary in response to the variations of streamflow. We

agree that such post-analysis is not necessary in this study and may even dilute the goal of this study; therefore, in the revised manuscript, we will replace this part with more meaningful comparative analyses of different feature importance methods.

We are sorry for the confusion due to some contradictory statements such as "the WFI has an advantage over PFI and MDI as it does not account for predictive accuracy so the risk of overfitting will be greatly reduced" and "the comparative study also shows that the predictors identified by WFI achieved the highest predictive accuracy on the testing dataset". In fact, we were trying to talk about two things. Since the importance scores based on WFI and MDI are calculated using the training dataset, the splitting actions without using predictive accuracy as criteria (i.e., the proposed splitting rule) may lead to less overfitted decision trees. Our results indicated that the less overfitted decision tree performed better over the testing dataset. In the revised manuscript, we will rewrite such contradictory statements to avoid confusion.

We much appreciate the referee's careful and insightful reviews, which are very valuable for us to improve the presentation of the proposed method. We will carefully review the papers listed and improve our manuscript.

Best regards,

Kailong Li, on behalf of the team of authors