Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2021-65-RC2, 2021

Comment on hess-2021-65

Anonymous Referee #2

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-RC2, 2021

Summary
This paper focuses on feature importance scores. These scores are widely computed in the (hydrological) literature to increase interpretability in machine learning applications. More precisely, the paper introduces a new feature importance method (and a new splitting rule), and applies this method to three interconnected irrigated watersheds. Comparisons with existing feature importance methods are also conducted for the same watersheds.

General comments
Overall, I believe that the paper is meaningful and interesting. Also, much work has been done for it.
However, I currently have several major comments that, to my view, should necessarily be addressed and, therefore, I recommend major revisions.
The key direction in which revisions should be made is the following: As the paper introduces a new splitting rule (that is of general use), it should be written accordingly and not as a work entirely placed in a hydrological context. More precisely, its introduction section should start by presenting detailed information on existing feature importance methods and splitting rules (thereby introducing the reader to the study's background), continue by stating what the new splitting rule offers compared to the existing ones (in summary), and lastly state that three hydrological applications are conducted. Most of the works cited in this review do likewise. Also, these applications should not be treated as if they consisted some type of proof, but as examples illustrating how the new method could be adopted in hydrology. A more appropriate proof that the method performs well compared to existing ones (e.g., an extended empirical investigation using a much larger dataset or a rigorous theoretical explanation) should also be provided, to my view.

Specific comments
1) To my view, the paper primarily focuses on a problem that is not hydrological −by its own nature− but algorithmic and more general, and it only secondarily presents a hydrological application. Therefore, it should be written accordingly and not as a work entirely placed in a hydrological context. In fact, the current version of the manuscript could confuse the reader, as it leaves the impression that the new method is motivated by

hydrological discussions (mostly those made around the equifinality principle), and not by the machine learning literature and the need to provide better feature importance methods (which are, of course, needed in hydrology, but not only in hydrology).
2) The motivation of the paper seems to be related to the equifinality principle-concept (which is extensively studied in hydrology). Nevertheless, it is unclear to me how the provision of better feature importance scores could solve the "equifinality problem". Further, I am not sure if equifinality could be referred to as a "problem", as it only implies that different modeling solutions could lead to outcomes of similar quality-value.
3) More generally, the concepts of equifinality, interpretability, collinearity and predictivity are explained, discussed and presented to be connected in a way that could confuse the reader. Thus, I think that the related background should be re-examined and that the related parts of the paper should be updated accordingly.
4) Key literature pieces on splitting rules and decision trees are also missing from the manuscript, despite the fact that they are necessary for covering the work's background.
5) Much effort has been put for placing this research into a hydrological framework. However, my general feeling is that there is something artificial about this, which does not even offer much to the paper, as it does not make the algorithmic-machine learning part smoother and easier to capture. To my view and given the main contribution of the paper (which is the introduction of a new feature importance method and a new splitting rule), more attention should be placed to which the existing feature importance methods are (e.g., the Gini, permutation, conditional permutation methods), which their theoretical properties are and what the new method does compared to them.
6) To my view, lines 249-258 are the most important part of the manuscript (that ideally should be extended and written rigorously), as they communicate some key advantages of the proposed method. However, it seems to me that these lines present only thoughts (and perhaps the overall rationale behind the method's conceptualization), and that they do not provide any proof, neither are they connected with other concepts discussed in the manuscript (e.g., equifinality, interpretability and more). Further, I am not sure that a real connection with these concepts exists.
7) More generally, the paper does not seem to follow one of the standard paths appearing in the literature for justifying the introduction of new algorithms in general and splitting rules for decision trees in particular. These paths are (a) the study of the method's asymptotic properties (i.e., an assessment through a theoretical investigation), and (b) empirical tests using large datasets (preferred when a theoretical investigation is too difficult). The three empirical examples currently presented in the paper do not provide either a theoretical or an empirical justification that the new method is well-designed (and, therefore, this justification is currently missing from the manuscript). Please note here that I do not mean to imply that the proposed method is worse than others. Instead, I think that further justifications are required at the moment for the presentation of the new method to become complete and for its properties to be understood.
8) Furthermore, it is unclear to me (if and) how the Bayesian model averaging method supports in a straightforward way the assessment of the new feature importance method. Also, XGboost does not seem to be (absolutely) necessary for reaching the paper's objectives (as the boosting and random forest algorithms differ a lot). A general feeling of mine (which might be due to missing justifications) is that some of the methodological pieces are only artificially connected with the rest.
9) Additionally, some contradictory statements exist throughout the manuscript making the latter a bit hard to follow. For instance, in the abstract it is written that "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", while later it is written that "the comparative study also shows that the predictors identified by WFI achieved the highest predictive accuracy on the testing dataset".
10) Papers that could be consulted for improving the presentation of the new method, in line with the above-provided comments, are listed here below (but many more exist, while an attentive literature review is currently missing from the paper):
a) Athey et al. (2019): This paper presents a splitting rule for maximizing heterogeneity.

b) Bénard et al. (2021): This paper provides better definitions and explanations of concepts like interpretability, simplicity and predictivity, and could help in putting the manuscript in a broader context, in the machine learning community.
c) Du et al. (2021): This paper presents another splitting rule.
d) Epifanio (2017): This paper presents an assessment of a new approach to assessing variable importance.
e) Friedberg et al. (2020): This paper presents another splitting rule.
f) Gregorutti et al. (2017): This paper investigates the relationship between correlation and permutation importance measures.
g) Ishwaran et al. (2008): This paper presents another splitting rule.
h) Roy and Larocque (2012): This paper presents another splitting rule.
i) Scornet (2020): This paper presents a theoretical investigation of the MDI. Also, it provides a clear discussion of the concept of interpretability.
j) Strobl et al. (2008): This paper presents the conditional variable importance metric, which is among the most popular and old variable importance metrics.
k) Wager and Athey (2018): This paper presents another splitting rule. It also introduces causality for random forests.

References

Athey, S., Tibshirani, J., Wager, S., 2019. Generalized random forests. Annals of Statistics, 47(2), 1148–1178. doi:10.1214/18-AOS1709.
Bénard, C., Biau, G., Veiga, S., Scornet, E., 2021. Interpretable random forests via rule extraction. International Conference on Artificial Intelligence and Statistics, pp. 937–945.
Du, Q., Biau, G., Petit, F., Porcher, R., 2021. Wasserstein Random Forests and Applications in Heterogeneous Treatment Effects. International Conference on Artificial Intelligence and Statistics, pp. 1729–1737.
Epifanio, I., 2017. Intervention in prediction measure: A new approach to assessing variable importance for random forests. BMC Bioinformatics, 18(1), 1–16. doi:10.1186/s12859-017-1650-8.
Friedberg, R., Tibshirani, J., Athey, S., Wager, S., 2020. Local Linear Forests, Journal of Computational and Graphical Statistics. doi:10.1080/10618600.2020.1831930.
Gregorutti, B., Michel, B., Saint-Pierre, P., 2017. Correlation and variable importance in random forests. Statistics and Computing 27, 659–678. doi:10.1007/s11222-016-9646-1.
Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S., 2008. Random survival forests. Annals of Applied Statistics, 2(3), 841–860. doi:10.1214/08-AOAS169.
Roy, M.H., Larocque, D., 2012. Robustness of random forests for regression. Journal of Nonparametric Statistics, 24(4), 993–1006. doi:10.1080/10485252.2012.715161.
Scornet, E., 2020. Trees, forests, and impurity-based variable importance. arXiv:2001.04295.
Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinformatics 9, 307 doi:10.1186/1471-2105-9-307.
Wager, S., Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113 (523), 1228–1242. doi:10.1080/01621459.2017.1319839.