Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., author comment AC1
https://doi.org/10.5194/hess-2021-481-AC1, 2021

Reply on RC1

Johannes Laimighofer et al.

Author comment on "Parsimonious statistical learning models for low-flow estimation" by Johannes Laimighofer et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-481-AC1, 2021

We want to thank Reviewer #1 for the comments and valuable suggestions on our manuscript. Below you find our response and the changes we implemented in the document.

Throughout the manuscript, the term parsimonious is used for characterizing the statistical learning models due to reducing the number of predictor variables to include the important ones. In statistical theory, the term parsimonious mainly refers to the number of parameters of the model, which is still large, e.g. in a random forest or boosting-based model with few predictor variables.

We agree that the term parsimonious can be misleading if compared over all models, as the degree of freedom will be larger for simpler models as LASSO compared to a non-linear boosting approach. However, we argue that using less variables in a random forest model still reduces the complexity of the random forest model. So producing parsimonious models refers to reducing the number of variables for each individual model in our case. To avoid a misunderstanding we changed the term parsimonious to more parsimonious where it suits.

It is self-contradicting to state that non-linear relationship can be captured by linear learning models (lines 14, 15). Actually, the term "non-linear relationship" cannot be defined. Furthermore, a more accurate classification of models considers their flexibility and interpretability, while the more flexible models can model better relationships, albeit this costs to their interpretability.

The term „non-linear relationship" in this sentence was misleading, we meant non-linear hydrological processes, not non-linear coefficients in a linear model. We changed the sentence on line 13-15 to:

Line 13: A direct comparison of linear and non-linear models reveals that non-linear relationships can be sufficiently captured by linear learning models, so there is no need to use more complex models or to add non-liner effects.

Changed to: A direct comparison of linear and non-linear models reveals that non-linear processes can be sufficiently captured by linear learning models, so there is no need to use more complex models or to add non-linear effects.

The classification "statistical" vs "machine" learning models is not clear either. For instance, Support Vector Regression or RF can be considered statistical learning models also.

We did not want to make a distinction between machine learning models and statistical learning models as most models are indeed the same and the two terms are often used synonymously. We used statistical learning as we are more used to the terms used in statistical theory than to machine learning terms (e.g. coefficients vs. weights, variables vs. features).
To avoid a misunderstanding, we added the following clarification to line 24:

Line 24: Regression methods cover a wide spectrum of models and especially in the last decade there was gaining interest in statistical learning models in hydrology (Abrahart et al., 2012; Dawson and Wilby, 2001; Nearing et al., 2021; Solomatine and Ostfeld, 2008).

Changed to: Regression methods cover a wide spectrum of models and especially in the last decade there was gaining interest in statistical learning models in hydrology (Abrahart et al., 2012; Dawson and Wilby, 2001; Nearing et al., 2021; Solomatine and Ostfeld, 2008), with the terms statistical learning and machine learning being used synonymously.

Additionally we changed Support Vector Machine Regression to Support Vector Regression throughout the manuscript.

Some claims related to the fact that studies mostly claim that non-linear models outperform linear ones could be relaxed. The literature includes studies claiming the opposite. Some of them can be found in the references list of the manuscript.

Line 45: Tree based methods performed better in terms of point prediction for the CAMELS data set (Tyralis et al., 2021) or an Australian data set of 605 stations (Zhang et al., 2018).

Changed to: Tree based methods performed better in terms of point prediction for the CAMELS data set (Tyralis et al., 2021) or an Australian data set of 605 stations (Zhang et al., 2018), but both studies showed good performance for less complex linear models.

Line 51: A general tendency visible from most studies is, that more complex models seem to perform better than more parsimonious ones, making model interpretation difficult and plausibility of parameters hard to judge.

Changed to: Results of Worland et al. (2018) and Ferreira et al. (2021) indicate that more complex models seem to perform better than more parsimonious ones, making model interpretation difficult and plausibility of parameters hard to judge.