Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2019-650-AC1, 2020

Hydrology and Earth System Sciences
Open Access
Discussions

Interactive comment on "Predicting tile drainage discharge using machine learning algorithms" *by* Saghar Khodadad Motarjemi et al.

Saghar Khodadad Motarjemi et al.

sa.m@agro.au.dk

Received and published: 21 February 2020

We would like to thank the referee for all the useful and constructive comments, which would enable us to improve the manuscript. Precision and the detailed assessment is very much appreciated. In the following, we provide replies to the referee's comments. We also describe the changes, which we will make in the final manuscript to accommodate the referee's comments.

1: Referee comment: However, the manuscript lacks more detailed explanation on motivation of such study and final conclusions on applicability of the results. The latter is most probably the case due to the miss-conception of the validation process. Finally, after performing the cross validation, the study does not extract any new knowledge,

rather discuss differences in cross-validation techniques - which clearly does not fit the scope of the journal.

Authors' reply: We agree that such explanation is missing in the current version of the manuscript and it would be included in the revision. The main objective of this project is to produce a map of yearly tile drainage discharge on a national scale. More than half of the agricultural fields in Denmark are artificially drained, and studying the pesticide and nutrient leaching through the tile drains has been a focus for many years. Such studies requires costly measurements of drainage discharge at different scales and a discharge map could be used as a fair proxy. On the other hand, mitigation techniques such as constructed wetlands have been widely used in the recent decade targeting the removal of excessive nutrients from the drainage water, before it reaches the fresh water reservoirs. Quantification of drainage discharge is of high importance for the design of the constructed wetlands and a national scale map could assist in this regard. The current study employed all the available data from the stations representing different geological regions with various characteristics to train models that can perform as predictive tools, applicable to the new datasets. The detailed analysis on the validation process is to stress the importance of resampling procedure and the spatial bias that it could cause in this type of data. Most of the 53 stations have multiple years of measured drainage discharge and unlike percolation, the other covariates are constant with time. Training the model on 90 % of the data increases the possibility of having the same station in the training dataset and in the test dataset. Leave-station-out guarantees that the target station does not also appear in the training dataset, however, it would still bias the accuracy assessment as it has similarities with neighboring stations. Which is why leave-cluster-out resampling is the least biased when training the model, as it excludes all the stations within 10 km (as a cluster) from the training dataset. Finally, the importance of resampling methods in the application of machine learning in hydrology has not been widely discussed (to our knowledge) and the findings of this study could be of relevance.

2: Referee comment: The study itself has been thought systematically on how to approach the modelling phase. However, some phases were misconducted. First of all, the time scale of the study is considered to be annual in regard with the output, which is not clearly specified how then the input has been encoded/aggregated, knowing the fact that meteo data are available on daily bases. Next problematic approach is using mechanistic models to encode/represent the input in the modelling process. Such case is with meteorological data that are run through water balanced model EVACROP.

Authors' reply: The percolation (Db) was calculated on a daily basis based on the precipitation and evapotranspiration (for winter wheat as the most dominantly cultivated crop in Denmark), and later was summed up for the entire hydrological year as a yearly value. The drainage discharge data was mostly available on a yearly basis and for those on a daily scale; same method was used to calculate the yearly values. The aforementioned information will be included in the revised manuscript. The importance of the Db calculations via the simple water balance model EVACROP will be discussed during further replies.

3: Referee comment: To this end, I would rather say that finding out that percolation (Db) is most important attribute upon running huge machinery is not an added value, as that fact is proven by theory and more specific by correlation of both variables discharge and percolation, which is obvious from Figure 7. Rather more interesting contribution would be to see which of precipitation and/or evapotranspiration is more significant in combination with different/specific landscape and soil characteristics. Similarly, second most importantly identified covariate - elevation - is pretty difficult to be simply explained as cause for discharge. The small range of values with pretty small sample size cause a behaviour as a clustering bias, especially if experimental sites are uniformly (equidistantly) distributed along the given range. So instead of discovering more interesting patterns, those are replaces with single covariate that encapsulate different processes under the hood. Therefore, I would rather see what is happening if this covariate is removed.

Authors' reply: The most important reason for using calculated percolation instead of measured precipitation is the summer precipitation events, which would not affect the drainage discharge. Denmark has a precipitation deficit in the summer, and the root zone therefore stores most of the precipitation during the summer. Precipitation is distributed over the entire year. Large precipitation events occur both during the winter half year as well as during the summer half year. In our study, we aim at predicting the yearly drain discharge due to the low temporal resolution of a large portion of our drainage discharge data set. Using the EVACROP-simulated Db instead of precipitation will eliminate drain discharge during the growing season where large precipitation events seldom will trig any significant tile drain discharge events. However, the model was also ran with the precipitation as a covariate instead of percolation and the results of model accuracies and the most important covariates are presented in Figures 1 and 2, respectively. The accuracy of the models are not significantly different compared to the results reported on the manuscript with Db as covariate. In Figure 7 of the first version of the manuscript, tile drainage discharge shows a U-shaped relationship with elevation. This shows that topography can have an effect on discharge, which Db does not account for. As we used the 1-dimensional EVACROP model to calculate Db, this covariate does not account for topography. The observed pattern is most likely a combination of several effects. Firstly, higher elevations receive more precipitation, which would increase discharge. Secondly, EVACROP does not account for surface flow. Lower elevations are likely to receive additional water from upslope positions, which would increase discharge. Thirdly, lower elevations will often have a shallower depth to the groundwater, and groundwater flow from higher positions may therefore contribute to the increased discharge. Together, these explanations show why intermediate elevations may have less discharge than higher and lower elevations. To assess the effect of excluding DEM, the model was run without the covariate and the results are shown in Figures 3 and 4. The accuracy of the models do not show a noticeable change after excluding the DEM as a covariate. Regarding the most important predictors, horizontal distance to the channel (Hdtochn) and clay content in the D horizon (Clay.D) appear as the second and third most important variables after precipitation. These two covariates also had high importance in the models that used elevation as a covariate. The results mainly show the adaptive behavior of machine learning models. When an important covariate is missing, the algorithms can to some extent use other correlated covariates to act as proxies. For example, valley depth and vertical distance to channel may act as proxies for elevation. Presented results in this discussion could be also included in the paper to provide a comparison.

4: Referee comment: Regarding the manuscript, the sections introduction and data are well described. The methodology and validation part is also fairly good described, except the part for how the importance of covariates is performed - especially knowing the fact that RF is not that open model so to be able to easily extract the most important covariates.

Authors' reply: We fully agree that important information about the importance measures of the predictors is missing on the manuscript and will be included after revision. We chose %IncMSE as the measure of variable importance in the RF model. The %IncMSE indicates the increase in the MSE of prediction, drainage discharge in this study, as a result of one variable being permuted. The higher the value of %IncMSE is, the more important the variable is for the regression of the RF model. For the Cubist model, each predictor had a value of the VarImp (%), which is a linear combination of the usage of each variable in the rule conditions and the linear regression models. We used this value to measure the importance of each predictor in the Cubist model. We calculated these two measures using the function varImp in R package caret, which we used for training the models.

5: Referee comment: Results and discussion section are lacking more details and focus on actual findings and less (or at least not that dominantly) on performance from different validation schemas. Validation schemas are well defined, and in discussion difference in performance of the models should be discussed - talking of which, spatial bias is not mention upon introduction. Such given discussion sounds more of evalu-

ating three different validation schemas, rather than discussion of new findings in the domain of hydrology.

Authors' reply: The recommendation will be taken into account for the revision of the manuscript and necessary changes mentioned by the referee will be implemented.

6: Referee comment: l.20: "This work opens up for a better understanding of the dynamics of tile drainage discharge and proves that machine-learning techniques can perform as predictive models in this specific concept." - too optimistic conclusion without good ground for such claim.

Authors' reply: We fully agree that the sentence is formulated in a wrong way. As the discharge is annual in different time periods for each station, the study cannot open up for better understanding of the "dynamics".

7: Referee comment: l.229: "The proposed tile-drainage discharge predictive model is not dependent on the climatic and constantly measured data and makes it possible to use different geographical properties as predictive parameters." - this is absolutely not true as percolation is derived from a model that uses at input precipitation and evapotranspiration data.

Authors' reply: We fully agree with the comment and the statement will be excluded from the manuscript.

Please also note the supplement to this comment:
https://www.hydrol-earth-syst-sci-discuss.net/hess-2019-650/hess-2019-650-AC1-supplement.pdf

---

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-650, 2020.

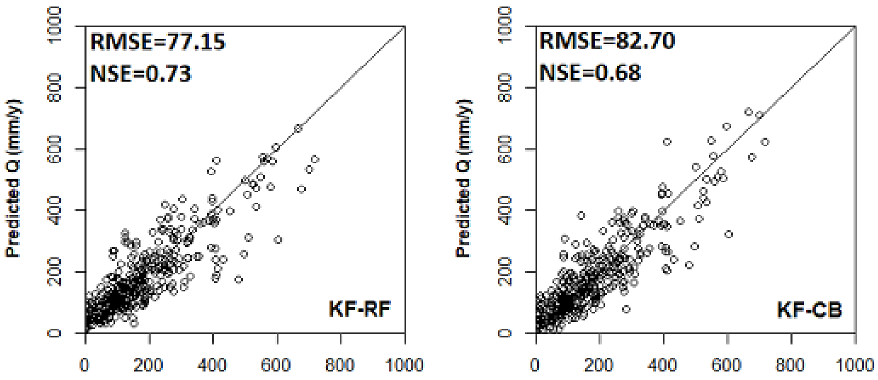

Fig. 1. Model performance of KF-RF: K-Fold cross-validated random forest model, KF-CB: k-fold cross-validated cubist model with precipitation (P) instead of percolation (Db).

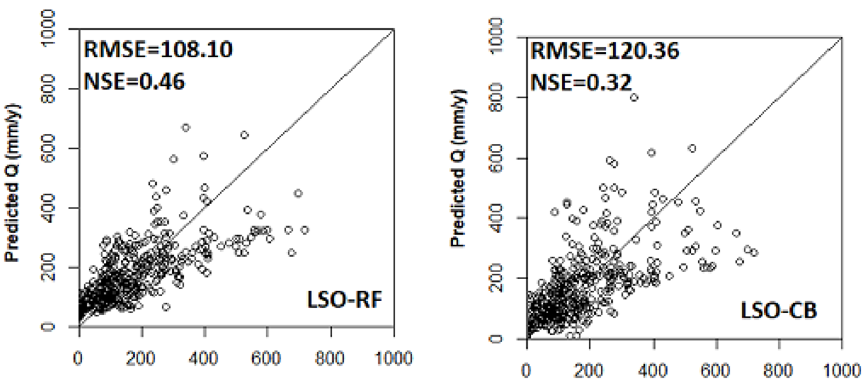

Fig. 2. Model performance of LSO-RF: Leave station out cross-validated random forest model, LSO-CB: Leave station out cross-validated cubist model with precipitation (P) instead of percolation (Db).

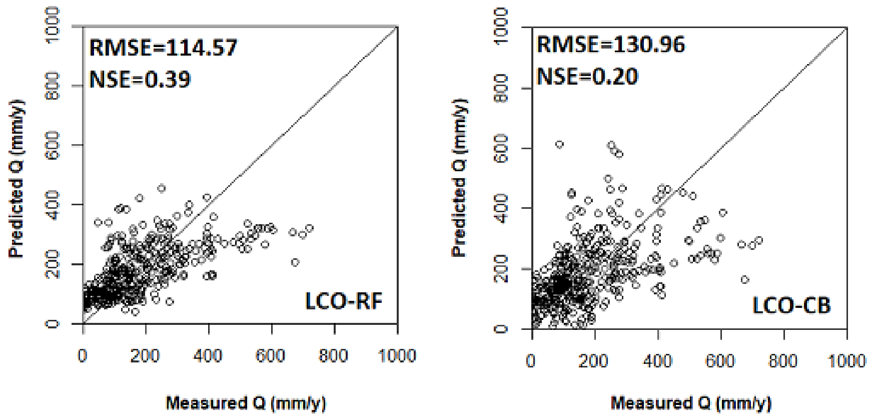

Fig. 3. Model performance of LCO-RF: Leave cluster out cross-validated random forest model, LCO-CB: Leave cluster out cross-validated cubist, with precipitation (P) instead of percolation (Db).

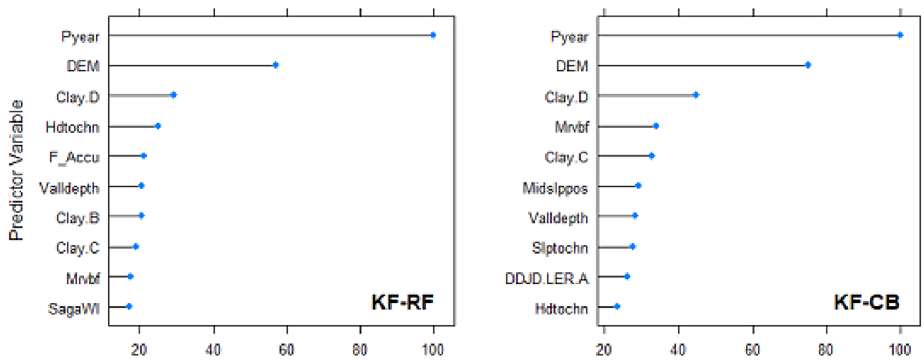

Fig. 4. Top 10 most important covariates of KF-RF: K-Fold cross-validated random forest model, KF-CB: k-fold cross-validated cubist model, with precipitation (P) instead of percolation (Db).

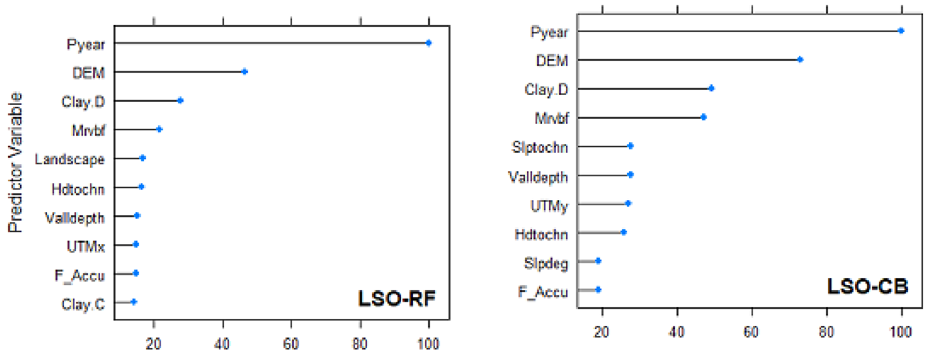

Fig. 5. Top 10 most important covariates of LSO-RF: Leave station out cross-validated random forest model, LSO-CB: Leave station out cross-validated cubist, with precipitation (P) instead of percolation (Db).

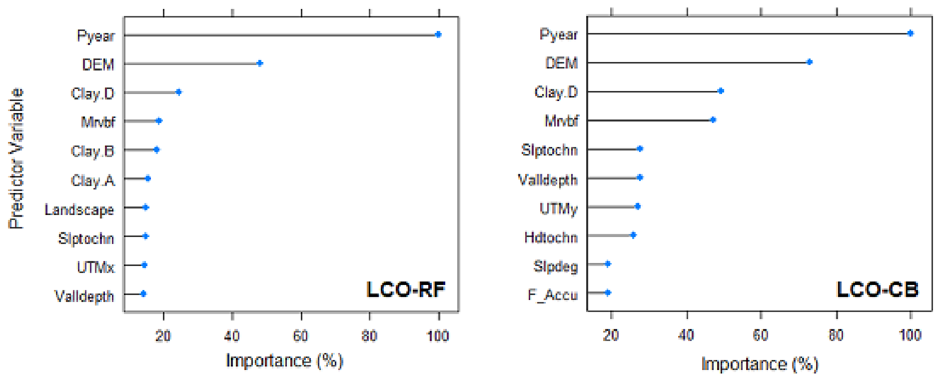

Fig. 6. Top 10 most important covariates of LCO-RF: Leave cluster out cross-validated random forest model, LCO-CB: Leave cluster out cross-validated cubist, with precipitation (P) instead of percolation (Db).

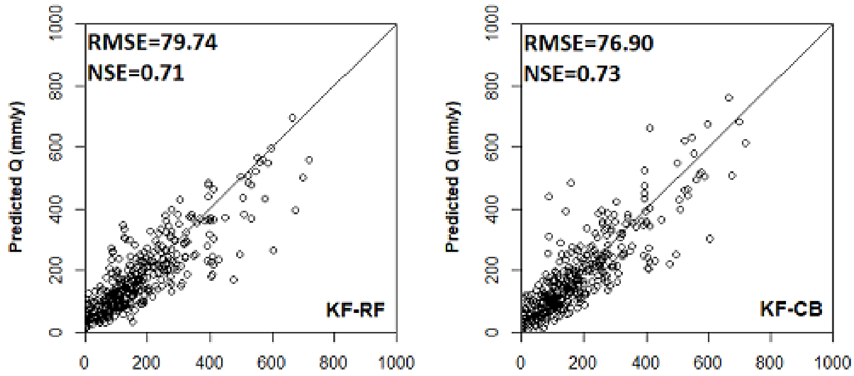

Fig. 7. Model performance of KF-RF: K-Fold cross-validated random forest model, KF-CB: k-fold cross-validated cubist model, when models was ran after excluding DEM.

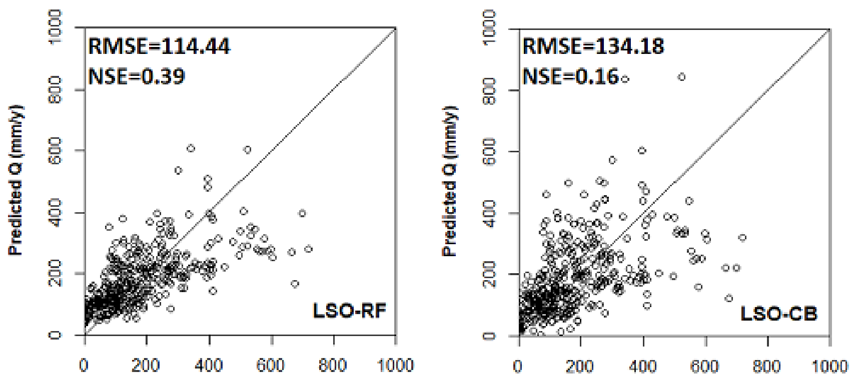

Fig. 8. Model performance of LSO-RF: Leave station out cross-validated random forest model, LSO-CB: Leave station out cross-validated cubist model, when models was ran after excluding DEM.

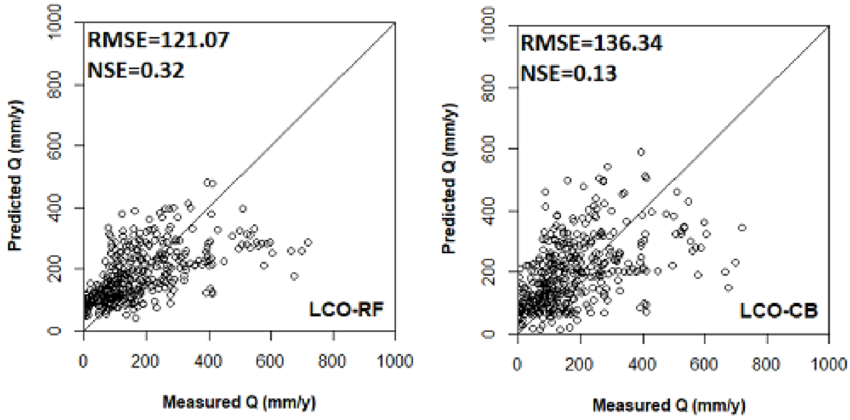

Fig. 9. Model performance of LCO-RF: Leave cluster out cross-validated random forest model, LCO-CB: Leave cluster out cross-validated cubist, when models was ran after excluding DEM.

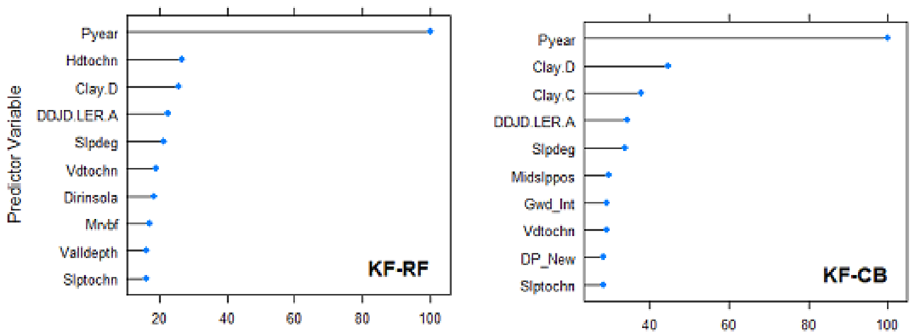

Fig. 10. Top 10 most important covariates of KF-RF: K-Fold cross-validated random forest model, KF-CB: k-fold cross-validated cubist model, when models was ran after excluding DEM.

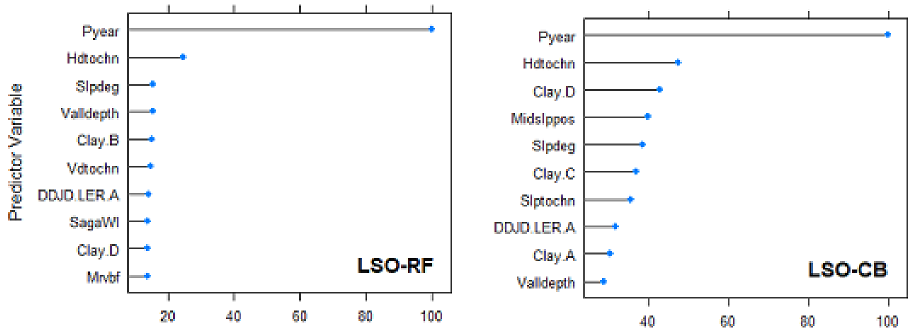

Fig. 11. Top 10 most important covariates of LSO-RF: Leave station out cross-validated random forest model, LSO-CB: Leave station out cross-validated cubist, when models was ran after excluding DEM.

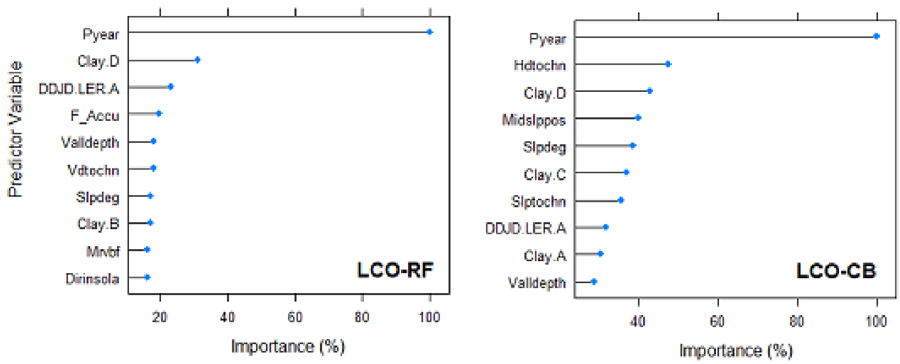

Fig. 12. Top 10 most important covariates of LCO-RF: Leave cluster out cross-validated random forest model, LCO-CB: Leave cluster out cross-validated cubist, when models was ran after excluding DEM.