Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2019-650-RC1, 2020

Hydrology and Earth System Sciences
Discussions

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

Anonymous Referee #1

Received and published: 23 January 2020

The manuscript entitled "Predicting tile drainage discharge using machine learning algorithms" is a well-written manuscript, the objectives are clearly stated, the Introduction section provides a sufficient description of the machine learning algorithms used in this study also to non-experts, and paragraphs are concise and logically connected. Although I appreciated the relatively short length of the manuscript, the methods need to be expanded and clarified. This research reports the application of machine learning to predict yearly discharge values in tile drains in Denmark. The Authors integrated a sufficient number of datasets from multiple sources to build the predictive model. However, the number of measurements reduces to 53 values and the datasets appear to be redundant rather than distinct. As a consequence, my concern is that the model may not be able to capture interaction effects and nonlinear dependencies as well as it

could not really learn more information from similar datasets and possibly made poor decisions in ranking the datasets.

I believe the research would benefit from a deeper analysis of the effects of using stations with different ranges of historical measurements and the significance of the datasets, which may altogether bias the model. Figures and Tables might be revised to increase the clarity, the level of detail, and to ease Readers' understanding of the possible research limitations. The manuscript can be interesting for the scientific community working on machine learning applied in hydrology. But at the present state, I would not recommend it for publication because it is not possible to understand the soundness of the methods used to train the model. I am willing to review a revised version of the manuscript.

Below I report my major concerns of the methodology.

Abstract

L20: I have to disagree that this work opens up for a better understanding of the dynamics of the drainage discharge. As I discuss later, the Authors use average values throughout the observation period, and therefore, dynamic effects are loss.

Section 2.2

The Authors report They used data from 53 stations; 18 stations collect data from 2012 to 2016 and 34 between 1971 and 2009. The number of stations does not sum up to 53. The problems are that the measurements cover different time periods. The model may have been trained using data from 2016 in one location and data from 1971 in another location. There are no information on discharge trends in the stations, which may be available at those locations with long observations. Anyway, yearly values were calculated for each location neglecting possible trends and variance in data measurements. Can the Authors find a methodology to use only stations that are consistent in time? There are no information on data quality, while possible data gaps may exist. If

this was the case, how were they filled?

May I please ask the Authors to add a reference to indicate where precipitation measurements and evapotranspiration values come from?

According to the Authors' hydrological model, percolation out of the root zone is calculated as the difference between precipitation and evapotranspiration. Here, there exist some assumptions which have not been stated. For example, is it valid to neglect irrigation from the model? Is it valid to assume that the crop-specific coefficient Kc=1 to calculate the actual evapotranspiration from the potential one?

L113: May I please ask the Authors to report the accuracies of the digital maps in Table 1? In this regard, the Authors comment at L237 that accuracy error of the digital maps may influence their importance as covariates. If the Authors know the accuracies, They could carry out a sensitivity analysis using the available standard deviation as prior information and assess the prediction outcomes.

Section 2.3

How did the Authors integrate numerical and categorical variables? What was the approach followed by the Authors to convert categorical variables to numerical ones? May the Authors discuss what are the implications of such integrations with respect to the final predictions?

L156: Can I please ask the Authors to state how They extracted the covariate importance? In general, which software and packages were used to carry out the study?

L169: The Authors report the possibility to use methods to determine the most effective parameters, thus opening the opportunity to reduce the number of covariates. Did the Authors try to rerun the machine learning using a subset of covariates?

Section 3.1

Please add a reference to Table 2 where the accuracies of the methods are reported.

L179: The cluster analysis was an interesting approach. However, the clusters were different in size. Was there a relationship between overall accuracy and number/location of the stations excluded from the training set?

Section 3.2

L193: Can I please ask the Authors to use one term either percolation or discharge out of the root zone, for clarity?

L211: Please use a new Section for the Discussion.

L225-L230: This paragraph seems crucial for the understanding of the predictions but it is difficult to follow. The Authors here discuss the implications of having time-series covering different time-periods. Because Their explanation is not clear, it is difficult to be convinced about Their interpretation.

L229: The Authors state that the model is not dependent on climatic forcing. However, this is not because the effects of precipitation and evapotranspiration are accounted for, which are climatic variables.

L241: How can the Readers know that the areas with high catchment area are the ones with larger Q/Db? The Authors may use Figure 3 to show such relation? Maybe They could add some text to report the area.

L248: While it is possible that low-elevated areas are the ones with higher Q, it is difficult to think that distance from groundwater table or the depth to sink (does this refer to the depth to tile drain?) are not significant. Have the Authors tried to remove the DEM as covariate and see how the other covariates rank?

L258: I have to disagree with the Reviewers statement: "the possible variations between years as well as outputs in relation to future climatic scenarios can be studied". In fact, there is no analysis with regards of time; yet, the Authors are somehow contradicting Themselves as They stated at L229 that the model does not depend on climatic forcing. Therefore, no studies in relation to future climate scenarios can be carried out.

As far as the Figures are concerned:

Figure 1

It can be more informative. It would be ideal to have an ID that identifies each location and link the spatial information with the corresponding metadata compiled in a large additional table.

It shows that areas are quite far from each other and may follow peculiar dynamics. This strengthen my concern that the number of yearly values and covariates may not suffice to highlight the functioning of the system. Showing how the location cluster affect the predictions may support the interpretation of the results.

Figure 2

I would suggest to first show the relationship between measured variables (i.e., Q as dependent variable and P and ET0 as independent). At a later stage I would show the relationship between measured Q and predicted Q. Finally, I would show the relationship between Q and significant covariates. As of now, Figure 2b is not informative. It shows the relationship between the Discharge out of the root zone (Db) and the Precipitation, which the latter was used to calculate Db.

Figure 3

It would be more meaningful if additional information were provided to understand for which conditions Q is greater than P (e.g., low-lying areas, etc. . .).

Figure 4

I like the idea of showing the maps because they report the gradient The Authors could ease the Readers if the locations were indicated in the maps. It might be valuable to create insets and show scatterplots between measured Q with each covariate reported in the map, with an errorbar to indicate the accuracy of the maps at the location.

Figure 5

This figure makes me wonder: why if I do not use 1 station in the training set (LSO), I have worse accuracy than when I do not use 10% of the stations in the training set (KF). I think the Authors very briefly discussed this at L225. But I would kindly ask Them to further clarify and expand their analysis on this. Was it about the time coverage? The location? The accuracy of the maps? I believe it is possible to disentangle this.

Figure 7

Panels b,d,f show elevation, which seems to have big range-discrete values. Can the Authors please explain?

Table 2. Is there a p-value? Such value could be used to decide which features are significantly relevant and control possible false discovery rates.

MINOR COMMENTS

L57: Can I please ask the Authors to state by how much was the improvement in terms of performance of the machine learning compared to physically based models to predict tile drainage discharge?

L73: Can the Authors please summarize the accuracy to the Readers?

L156: Note that, the Authors are not using 6 models, but 2 models and validating each with 3 different methods.

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