Reply to the review from Referee #1:

We greatly appreciate the time and efforts of the referee to provide us with such a comprehensive and detailed review. Undoubtedly, the review will enable us to improve the manuscript.

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:

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.

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". However, we use 414 observation points, which are the total yearly discharge values and not the average. Since the temporal resolution of our data set only is on a yearly level the dynamic of the discharge relates to yearly differences and difference between catchments.

2:

Referee comment:

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.

Authors' reply:

We appreciate the precision on details and the number of older stations must be corrected to 35.

3:

Referee comment:

Section 2.2

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?

Authors' reply:

We agree that having the same time period for all the stations would have been ideal and we assessed different scenarios regarding this issue, however, one of the main objectives of this project was to have national scale predictions that will enable us to extract a drainage map. In order to have at least one station at the main geological regions in Denmark, we decided to use all the historical data from the available stations. Taking into account the recommendation, we would include (at least some examples) of the discharge trends in the long-term running stations on the revised manuscript.

All the meteorological data were measured either at the stations or the nearby stations.

4:

Referee comment:

Section 2.2

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?

Authors' reply:

None of the catchments/fields used in this study were irrigated. The most commonly sown crop in Denmark is winter wheat and the calculations of evapotranspiration were made accordingly.

The relevant information is missing on the current version of the manuscript and would be included during the revision.

5:

Referee comment:

Section 2.2

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.

Authors' reply:

The accuracy of the maps are reported in the references Adhikari et al., 2013 and Møller et al., 2018. The statement on L237 is missing the relevant citation to Adhikari et al., 2013 and will be included in the revised manuscript. Although carrying out a sensitivity analysis and assessing the prediction outcomes could be very interesting, the only map with high prediction error reported by the author (Adhikari et al., 2013) is the clay content map, which makes such analysis less necessary in the current manuscript.

6:

Referee comment:

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?

Authors' reply:

Both Cubist and Random Forest work by recurrently splitting the dataset. In this process, they can use categorical covariates as they are, and we therefore did not convert the categorical variables to numeric variables. This is the standard approach in the use of these two algorithms.

7:

Referee comment:

Section 2.3

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?

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. However, the model was not ran with reduced number of covariates as it would not have an effect.

We calculated these two measures using the function varImp in R package caret, which we used for training the models.

8:

Referee comment:

Section 3.1

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

Authors' reply:

In L174 the accuracies with the reference to the table 2 are reported.

9:

Referee comment:

Section 3.1

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?

Authors' reply:

For Random Forest, there was no correlation between the number of observations in the clusters and the accuracy obtained, both for RMSE and R2 ($p > 0.05$). For Cubist, there was no correlation between RMSE and the number of observations ($p > 0.05$), but R2 was correlated to the number of excluded observations ($R = -0.53$, $p < 0.05$). However, this finding is most likely due to the fact that it easier to obtain a high R2 with a small number of points. We will therefore not include this finding in the final version of the manuscript.

10:

Referee comment:

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?

Authors' reply:

Using two terms was an error and we would make sure to only use "percolation".

11:

Referee comment:

Section 3.2

L211: Please use a new Section for the Discussion.

Authors' reply:

A new section will be created for the discussion in the revised manuscript.

12:

Referee comment:

Section 3.2

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.

Authors' reply:

Re-formulation of L225-L229: Highest accuracies were achieved by KF cross-validated RF and CB. However, 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.

L229-L230: a wrong statement by authors, which will be excluded from the manuscript after revision. The models are still dependant on the hydrological data as the percolation (Db) is calculated based on climatic and constantly measured data.

13:

Referee comment:

Section 3.2

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.

Authors' reply:

In L102 and L108-L112 we give the relevant information. If it does not suffice more explanation will be included in the revised manuscript.

14:

Referee comment:

Section 3.2

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?

Authors' reply:

We appreciate the exactness of the Referee and it helped us to notice that depth to sink (BS), which is actually irrelevant for the drained areas, should be excluded from the covariates.

To assess the effect of excluding DEM, the model was run without the covariate and the results are shown in Figures 1 and 2. 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.

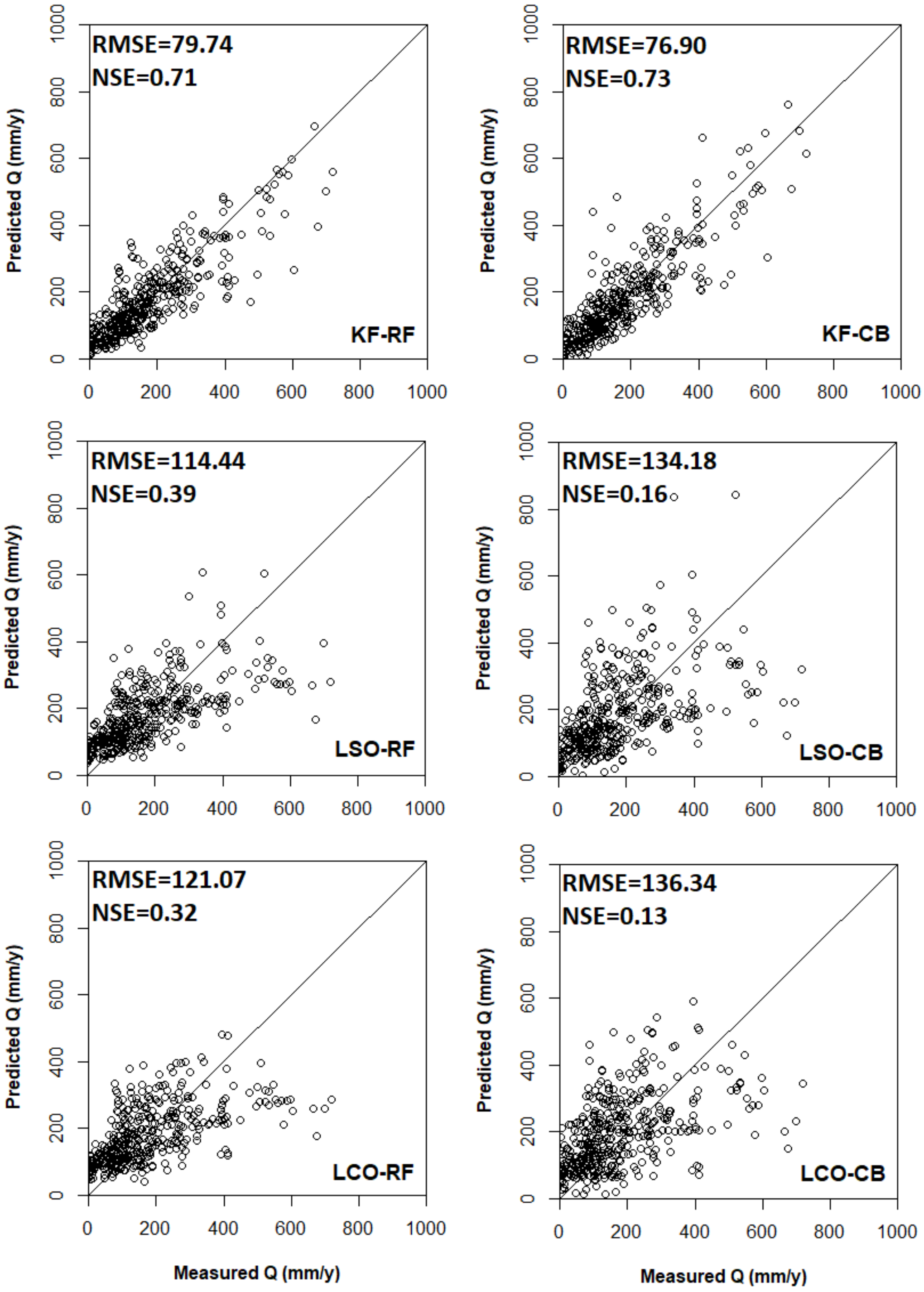

Figure 1. Model performance of KF-RF: K-Fold cross-validated random forest model, KF-CB: k-fold cross-validated Cubist model, LSO-RF: Leave station out cross-validated random forest model, LSO-CB: Leave station out cross-validated cubist model, LCO-CB: Leave cluster out cross-validated cubist model, when model was ran after excluding DEM.

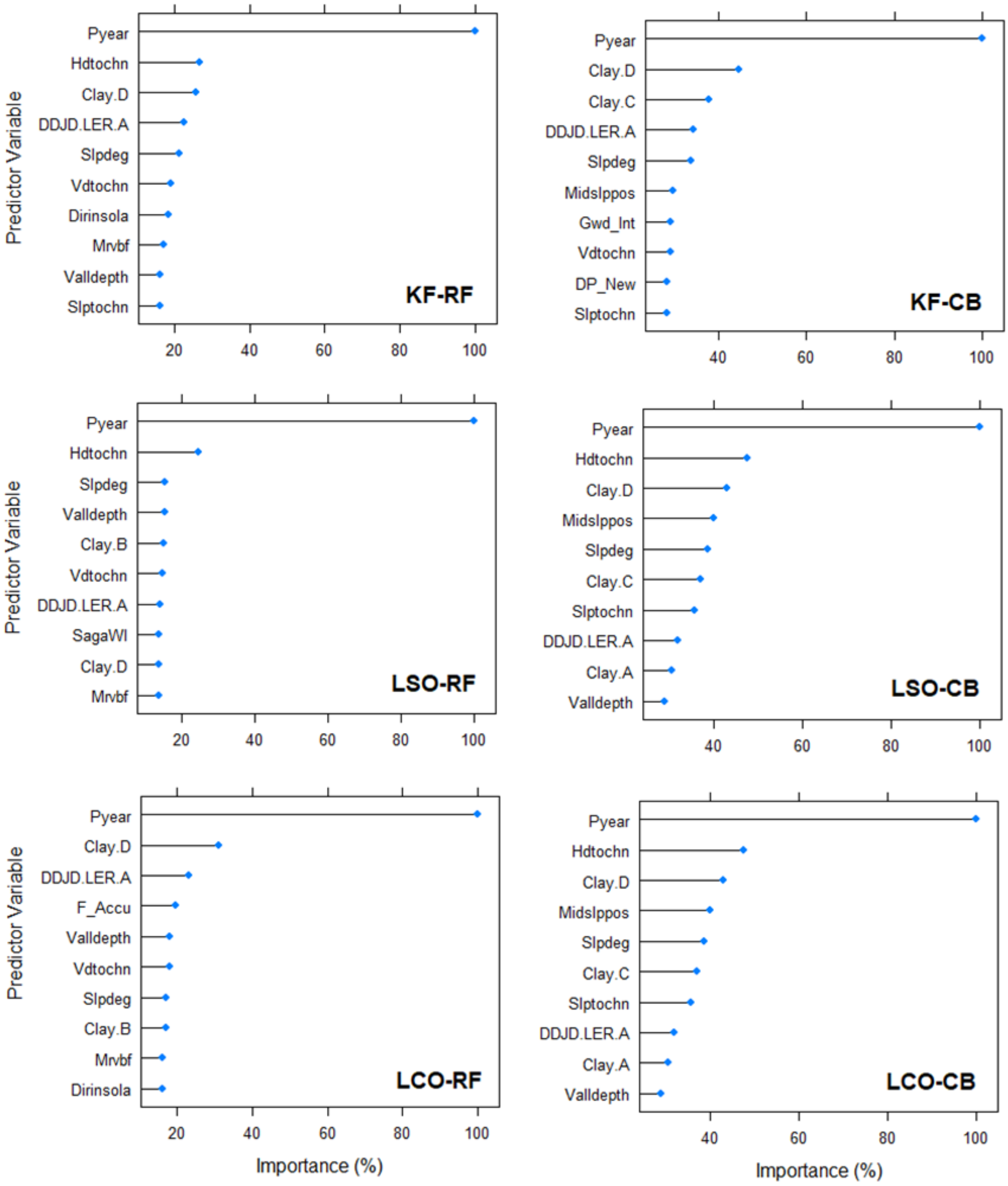

Figure 2. Top 10 most important covariates for KF-RF: K-Fold cross-validated random forest model, KF-CB: k-fold cross-validated Cubist model, LSO-RF: Leave station out cross-validated random forest model, LSO-CB: Leave station out cross-validated cubist model, LCO-CB: Leave cluster out cross-validated cubist model, when model was ran after excluding DEM.

15:

Referee comment:

Section 3.2

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.

Authors’ reply:

As mentioned earlier, we fully agree and a false statement by the authors will be corrected after revision.

16:

Referee comment:

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.

Authors’ reply:

Suggestion will be taken into account for the revision of the manuscript.

17:

Referee comment:

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.

Authors' reply:

We appreciate the exactness of the comment and the changes will be applied accordingly. We attached a figure demonstrating the relationship between P and Q.

18:

Referee comment:

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...).

Authors' reply:

The relevant information will be provided in the revised manuscript.

19:

Referee comment:

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.

Authors' reply:

The suggestion will be considered for the revised manuscript.

20:

Referee comment:

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.

Authors' reply:

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 (KF) increases the possibility of having the same station in the training dataset and in the test dataset. On the other hand, Leave-station-out (LSO) 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.

21:

Referee comment:

Figure 7

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

Authors' reply:

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.

22:

Referee comment:

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.

Authors' reply:

Could we kindly ask the referee to please elaborate a bit on this comment and question?

23:

Referee comment:

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.

Authors' reply:

L57: In the study of Kuzmanovski et al. (2015) the comparison showed overall improvement in the prediction of discharge through sub-surface drainage systems, and partial improvement in the prediction of the surface runoff, in years with intensive rainfall.

L73: The results of the study by Noi et al. (2017), showed that very high accuracy of Ta estimation (R2 > 0.93/0.80/0.89 and RMSE ~1.5/2.0/1.6 ◦C of Ta-max, Ta-min, and Ta-mean, respectively) could be achieved with a simple combination of four LST data, elevation, and Julian day data using a suitable algorithm. The summary of the accuracies for this study will be also added to the manuscript.

L156: We agree that stating six “models” might cause misunderstandings and we would try to explain better in the revised manuscript. We use two different algorithms and we train six models with different resampling methods.