Manuscript **hess-2019-212**: "Modelling of the shallow water table at high spatial resolution using Random Forests."

Correspondence to Julian Koch (juko@geus.dk)

Author response to Katherine Ransom. Reviewer evaluation in italic. Author reply in blue font.

General Comments

Overall this paper is well written, the methods are scientifically sound, and the work provides a substantial contribution to the current body of knowledge. The sensitivity analysis to provide local variable importance is highly useful and I am not aware of any other studies that provide such a map. This paper is suitable for publication in HESS. I have several comments, detailed below, that relate mainly to the methods descriptions that the authors can address mostly by providing more clarity or discussion related to the specific concerns.

Reply: We would like to thank Katherine Ransom for her thorough revision of our manuscript. We are very pleased that our modelling approach of the shallow groundwater system was generally well received. The comments made by Katherine Ransom raise valuable points and a rigorous revision following her suggestions will certainly improve the scientific quality of our work. We hope that the suggested changes will be appreciated and that the re-submitted manuscript will be rated fit for publication in HESS.

Specific Comments

In the data section, it is stated that 1,900 additional data points were used in the training dataset to represent areas where depth to groundwater is 0. However, later on, namely Figure 1 caption and in the Results section, it is unclear if the 15,000 additional points were used or if it was still just the 1,900. The data section states the data density of the additional points is the same as that of the measured data but this can't be the case if the authors only used 1,900 additional points. Please clarify throughout the text.

Reply: A total of 15k additional observations were placed along streams, coastline and in lakes. Our intention was to constrain the RF model with critical information that was otherwise not provided by the wells alone. We realized that including all 15k in the training would negatively affect the RF prediction, as the model was strongly biased to depth 0 m. Therefore we decided to use only a subset of the additional observations. We found 1900 a suitable number, because this reflects the same well density (1 well per km^2) as found in the well dataset. For the additional observations, the density refers to the area of surface water (stream, lakes and coastline) in 50 m grid resolution. In this way, the amount of wells and additional observations was balanced. However, in the residual kriging we used all 15k wells to correct the water table at locations with surface water where we expect a depth of 0 m. This will be stated more precisely in the revised manuscript and the caption of figure 1 will be changed accordingly.

In Section 2.2 how is the vertical distance to the nearest water body measured? Are the depth to water measurements involved in this calculation?

Reply: The vertical distance to the nearest waterbody is only in relation to surface waterbodies. For a given grid, we first find the nearest waterbody (lake, river, coastline). Then we compute the elevation difference of the given grid and the closest waterbody grid. This is will be stated more clearly in the revised manuscript.

Section 2.4 might be more appropriately labeled "Covariate Importance" or "Random Forest Sensitivity to Covariates"

Reply: We agree the term "Covariate Importance" would be more fitting to the standard RF terminology. However, the hydrological community may relate more to the term "Sensitivity", but readers may think of model parameter sensitivity which is not what we are addressing here. We will label the section “Covariate Sensitivity” instead.

I agree with the previous referee that the RMSE metric is probably better than R2 to quantify the covariate importance in the sensitivity analysis. Please discuss the reason to use R2 and the possibility to recalculate the sensitivity using RMSE.

Reply: We completely agree to the relevance of this point. In order to address this issue we computed covariate importance based on the relative increase in RMSE. The two figures below show the results based on the original assessment of the R^2 and the newly RMSE assessment. Although the percentages vary between the two metrics, the same conclusions in terms of relative covariate importance can be drawn. Therefore, we decided not to add the RMSE based figure in the revised manuscript, but instead, we will mention that we have conducted the sensitivity analysis based on RMSE and that it yielded the same conclusions in terms of covariate sensitivity ranking.

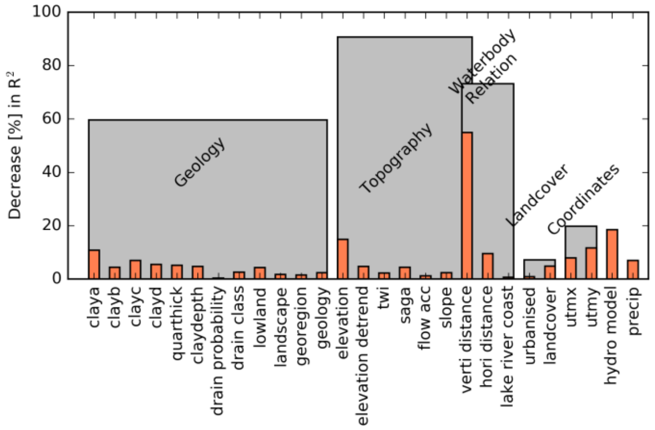

Figure 1 Covariate importance based on R2. Same as Figure 5 in the manuscript.

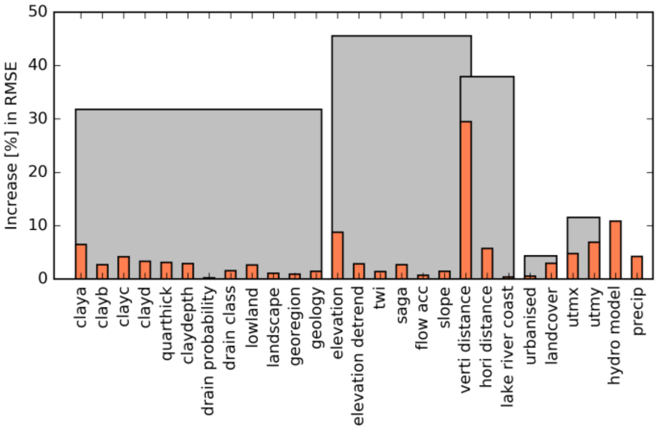

Figure 2 Covariate importance based on RMSE

In terms of the spatial mapping of covariance importance we already use the absolute difference between the original prediction and the permuted predictions. We have tested the squared differences but could not notice a significant change. We will be more explicit in the revised manuscript with respect to what metrics are used to compute the covariance importance and mention our tests of using alternative metrics.

It is unclear what the authors are referring to in Section 2.4 when they say “each simulation grid”. Do they mean each grid cell? The authors state: “prediction is repeated n times until the mean difference across n permutations converges for each simulation *grid." Do they mean the mean difference for each grid cell or the mean difference among all grid cells? Please clarify throughout the text.*

Reply: Correct, we meant each grid cell. For each grid cell, the difference to the original prediction is recorded for n permutations. We then compute the cumulative mean across these permutations and check if the mean converges.

Section 2.6 should include a description of the software used to calculate the QRFs. Was a special Python package available or was it programmed by the authors following the methods in Meinshausen, 2006?

Reply: We have used the scikit learn implementation of RF in python for our modelling work. To our knowledge, QRF is not implemented in scikit learn yet. Therefore, we have built our own QRF implementation as a workaround using the functions provided by scikit learn.

Section 2.6. This section seems incomplete. Please provide discussion on why the approach can be used if the underlying assumption of no covariance is violated and/or why the approach was used here. What is the purpose of the error propagation/how did the authors use it here? The explanation is provided on page 16 lines 10-14, but should be provided in the methods.

Reply: We acknowledge that the uncertainty propagation of RFRK and QRF was quite a leap given the current analysis in our manuscript. More work will be required to test the underlying assumption that the uncertainty sources have no significant covariance and thus can be combined. We have decided to remove section 2.7 and related text from the discussion. Figure 8 will be updated as well. However, we do believe it is a valuable contribution to present, compare and discuss the RFRK and QRF based uncertainty estimations.

In section 3.1 Random Forest Model, the authors state that "After initial testing, the RF model was parametrized as follows; the number of decision trees was set to 1,000, bootstrapping with replacement was applied to sample the training data, 33% of the covariates were considered to identify the optimal data split" and I am curious what the initial testing entailed and if the authors performed any tuning of these parameters, such as with a cross validation. It could be useful for the authors to more thoroughly describe the process and metrics used for selecting the number of trees and the percent of covariates selected for each tree. This description might also be more appropriate in the methods section.

Reply: With regard to the tuning of the RF hyper-parameters we have assessed two parameters in more detail: the number of tress and the n_leaf parameter, which controls the pruning of the decision trees. This initial test was conducted for a subdomain, which covers approximately 10% of our entire domain. Figure 3 and figure 4 show the RF performance for numerous combinations of n_tree and n_leaf parameters. We concluded that the performance converged for 1000 trees and that the trees should be fully expanded (n_leaf=1). Originally, we did not test the max_feature parameter, which controls how many covariates are selected randomly for optimizing each split. We chose 33%, because a lower number generally decreases computational time and increases diversity among the trees. Both aspects are desirable. For the purpose of this review, we briefly tested the sensitivity of the max_features parameter and observed that the R^2 for the oob prediction was affected in the third digit and the MAE (mean absolute error) in the second. Thus we conclude that the max_feature parameters is not sensitive for our application. We do not think that this information is necessarily relevant to readers and will therefore not expand the method description of the revised manuscript.

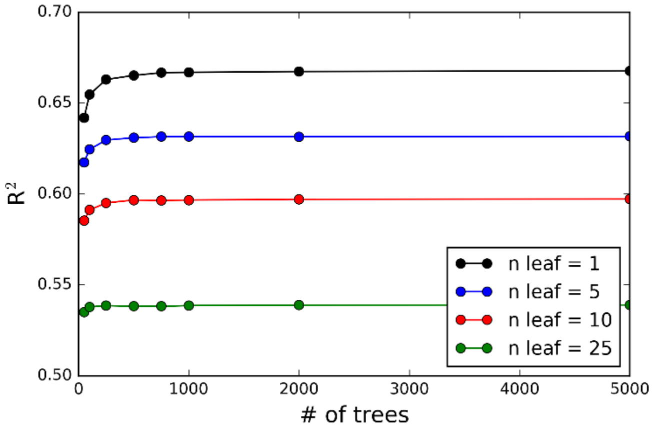

Figure 3 Initial testing of the RF hyper-parameters n_leaf and n_trees with respect to the R2 metric.

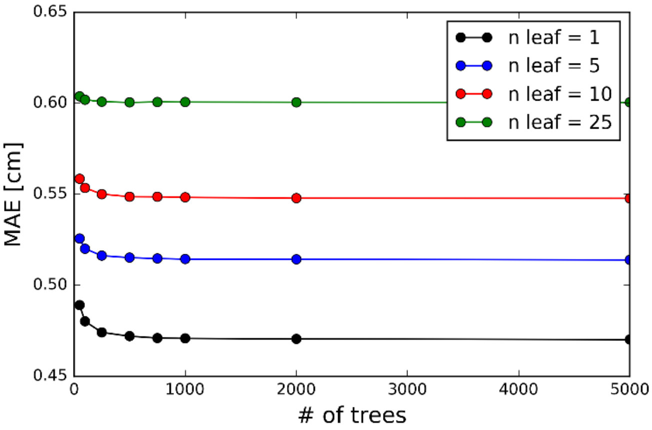

Figure 4 Initial testing of the RF hyper-parameters n_leaf and n_trees with respect to the MAE (mean absolute error) metric.

In section 3.1 Random Forest Model, the authors state that "The oob prediction can be considered as an independent validation test" and the authors did elaborate on this at the end of section 2.3. But readers may benefit from a reminder here that the contribution to the overall oob error from each observation is calculated based upon only the trees which did not contain that specific observation in the bootstrap and provide the reference (Breiman, 2001?). Though, I am not sure if I agree that the oob error can be used as an independent assessment of the generalization/validation error if this is what the authors meant. When predictions are made to unsampled areas or to unseen data, all 1000 trees are used. However, if the above is correct, the oob error is calculated for each observation based upon only a subset of the 1000 trees (n = 340), so the entire model is not assessed when calculating the oob error. The authors might want to consider calculating the testing error to a separate validation/testing set and comparing it to the oob error or providing more discussion on why the oob error also adequately quantifies the generalization error. Additionally, was the coefficient of determination a Pearson correlation coefficient or Nash-Sutcliffe? From the text I gather it is a Nash-Sutcliffe, this should be specified in the text.

Please provide summary statistics for the training data so readers can better understand the reported oob MAE and RMSE.

Reply: As suggested by the reviewer we will add further elaboration on how the oob predication is calculated to the revised manuscript. In order to clarify, it is not correct that only a subset of the trees are validated when using the oob prediction. The idea is that each tree uses its own bootstrap sample for traning.

In that way each tree also has its own oob sample that can be used to validated that specific tree. In the end, we can average over all trees where a sample has been retained as oob to obtain the final oob prediction. In this way, all trees have been validated when using the oob prediction.

We have not used the NSE metric to quantify model performance in the original submission. Instead we have applied the coefficient of determination (R^2), mean-absolute-error (MAE) and root-mean-squared-error (RMSE). In the revised manuscript, we will state the evaluation metrics more explicitly, but we will omit equations as these are quite generic metrics that the readers of HESS will be familiar with.

In order to investigate if the oob prediction is a reliable source to quantify the generalizability of a RF model we have conducted a 10-fold cross validation test. For this, the dataset was randomly split in 10 sets of approximately the same size. Then 10 RF models were trained on 90% of the data so that each set was left out once and could be used for validation purposes. The results are strikingly similar as compared to the oob prediction as shown in figure 5. Figure 6 depicts a scatterplot of the ~17k training samples comparing the predictions form the oob approach and the cross validation. In our opinion, the agreement is convincing which qualifies the oob prediction as an appropriate way to quantify the generalization error of our RF model. Furthermore, the statistics were very similar as well, as reported in the table below. We believe that this table and the addition of the 10-fold cv test is valuable for the revised manuscript and it will therefore be added to the re-submission. Please be aware that the oob metric scores vary slightly to the ones mentioned in the original manuscript. We regret that the script used to compute the scores did not read the latest data.

	R^2	RMSE m	MAE m
oob	0.56	1.13	0.76
10-fold cv	0.55	1.15	0.77

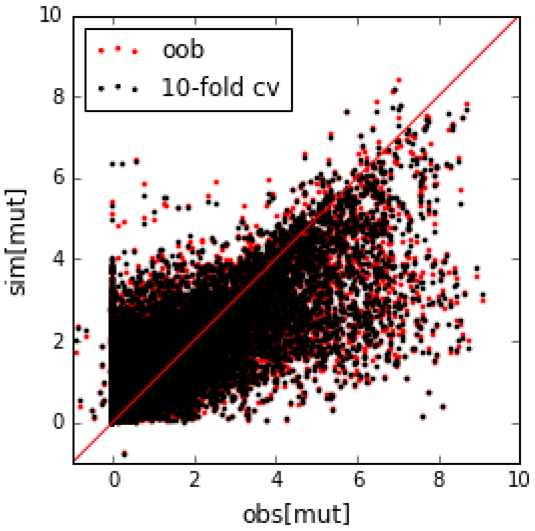

Figure 5 Comparison of oob prediction and 10-fold cross validation with respect to observations.

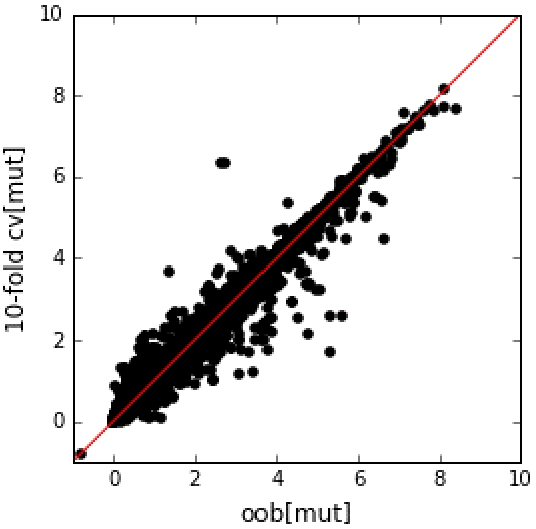

Figure 6 Comparison of oob prediction and 10-fold cross validation.

In section 3.1 and Figure 3, are the very shallow water table points which were consistently over-predicted the same additional points that were added (with 0 depth to water)?

Reply: This is not the case. The systematically overestimated shallow observations are consistently placed in the glacial till. It seems that the RF lacks covariate information to adequately reflect such conditions. Glacial tills can be very complex and at the current stage we do not have the required hydrogeological data with a relevant spatial resolution to resolve this issue. We will further elaborate on this shortcoming in the revised manuscripts.

Section 3.2 discusses the results of the prediction sensitivity analysis. From Figure 6 it does appear that this analysis was done on the grid cell level but please clarify in the text (see above).

Reply: Correct, the results of the sensitivity analysis that are presented in figure 6 (in the manuscript) are calculated on grid cell level. However, please be aware that we have implemented two approaches to quantify covariate importance. The first is the conventional assessment that applies the concept of permutation accuracy on the training dataset. Results for this analysis are given in Figure 5 (in the manuscript). The results of our novel contribution to assess sensitivity of the predication dataset at grid level are given in Figure 6 (in the manuscript). Both approaches are introduced in section 2.4. We will be more explicit about this distinction in the revised manuscript.

Section 3.3 should describe why all data including data not in the model was used for RFRK.

Reply: As mentioned earlier, the 15k additional observations were reduced to 1900 in order to be in balance with the well observations. However, this reduction was only affecting the RF training. For the RFRK we chose to include all additional observations to ensure that the final groundwater estimates are close to the surface at locations where surface water is present. This ensures physical consistency. The correlation length of the variogram model used to model the RF residuals is set to 200 m. This limits the effect of the additional observations with a groundwater depth of 0 m to a close vicinity.

From Figure 8 it is hard to tell if there is any variation among grid cells not located at a surface water location. Could the color scale be adjusted to better display the local variation for the RFRK?

Reply: It is correct that the uncertainty of the RFRK model does not vary at grids further away than 200 m (correlation length of the variogram) from an observations. Beyond 200 m distance the uncertainty will be equal to the sill of the variogram model (1.02 m^2). In that way changing the color scale would not change the figure. This is a natural characteristics of kriging based interpolations and will be mentioned explicitly in the revised manuscript.

Section 4.1. Did the authors compare model results with and without the additional data points of 0 depth to water? If such a scenario was tested it might be useful to discuss here.

Reply: We have compiled the figure below (figure 7) to address the effect of adding the additional observations to the RF model. The zoom extent is approximately 5 km from left to right and contains a lake, river system and wetlands. If we leave out the additional observations in the training dataset the final RF prediction does not capture the interaction between surface water and groundwater very well. In Denmark, it can be assumed that all surface waterbodies are connected to the shallow groundwater system. This example should underlie the importance of the additional observations in the applied RF model and will be discussed in further detail in the revised manuscript.

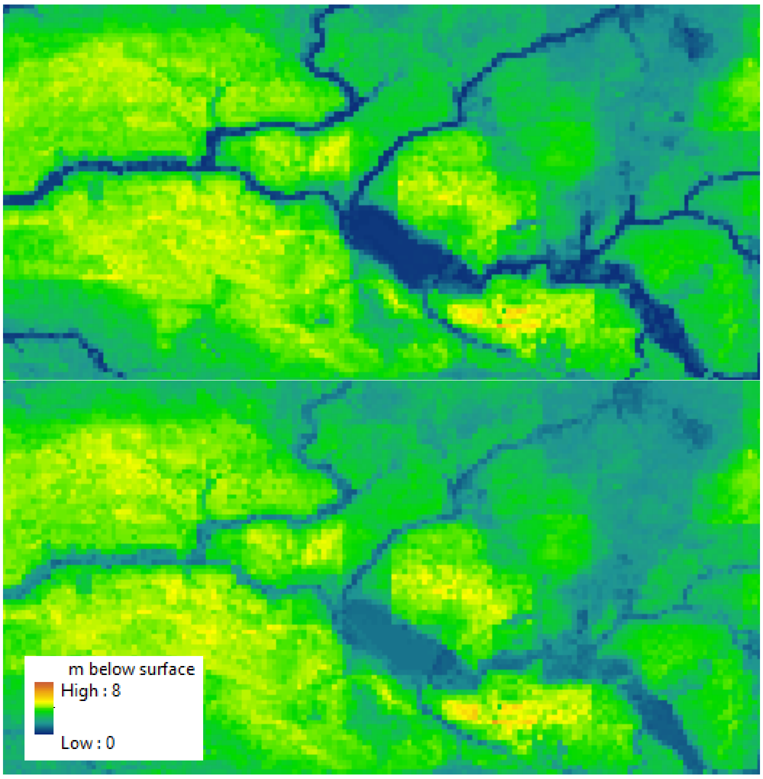

Figure 7 Results from two training scenarios: The top was trained against well observations plus the 1900 additional observation with a groundwater depth of 0 m. The bottom was trained exclusively against the well observations.

Section 4.2 Line 19-23 Were the covariates with low importance expected to be important relative to the covariates ranked as highly important? In addition to the possibilities the authors discuss, the drainage characteristics and topographic wetness index may also be overshadowed by the highly ranked covariates and could become important if the other covariates were removed from the model. If the RF model is not selecting the drainage characteristics and topographic wetness index covariates for splits very often or if splits on these variables occur far down in the trees (near the leaves) then we would not expect the permutations to be highly impactful. Along these lines, did the authors consider calculating other forms of variable importance such as relative importance based on reduction of RMSE attributed to each covariate within the model?

Reply: In one of our previous replies we showed the results of the covariate importance analysis based on the relative increase in RMSE as an addition to the R^2 based assessment and argued why it does not provide any additional insights. The statement by the reviewer is correct and gives a good technical explanation of why some covariates receive a low importance score. We will include some of these thoughts in the revised manuscript.

Technical Corrections

Reply: We appreciate these technical/editorial suggestions. We agree to all points raised by K. Ransom and will update the revised manuscript accordingly.

Table 1, Column 2, Row 9: “and” instead of “an”?
Figure 5 should have more descriptive labels for covariates, like Table 1.
Page 16 Line 8: do the authors mean each grid cell?
Page 17 Line 22: incomplete sentence?
Page 18 Line 11: “located” instead of “placed”