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 Anders Bjørn Møller. Reviewer evaluation in italic. Author reply in blue font.

General comments
I have read the manuscript "Modelling of the shallow water table at high spatial resolution using Random Forests" submitted to HESS by Koch et al., in order to provide a referee comment.
The manuscript is well structured, clear, concise and well written. It addresses the depth to the shallow water table, which is a highly relevant issue, and introduces a number of novel methods in doing so. Some parts of the introduced methods have great potential, not only for hydrological applications but for spatial predictions with machine learning in general.
My main concerns with the manuscript lie with some of the specific choices that the authors make in implementing the methods, especially related to the assessment of the accuracy and uncertainties of the predictions. I will elaborate on these concerns in the following section. However, given that the authors address them, the manuscript is highly suitable for publication in HESS.

Reply: We would like to thank Anders Bjørn Møller for his comprehensive review, which raises very thoughtful comments on our manuscript. We are very pleased to have received an overall positive evaluation of our manuscript and will gladly revise it following his comments to further strengthen the scientific quality of our work. We hope that the suggested changes, which are outlined in the response below, will be well received and that the re-submission will be regarded a significant contribution to HESS.

Specific comments
Firstly, I am wondering why the authors choose to map the depth to the shallow water table rather than the elevation of the shallow water table. I would expect the elevation of the shallow water table to show less spatial variation than the depth from the surface. It should therefore be easier to predict, all other things equal. I am sure the authors have good reasons for this choice, but I would like to see them stated explicitly.

Reply: The variable required by the stakeholders is the depth to the groundwater table, which is straightforward to interpret in the context of infrastructure planning and implementing of climate change adaption strategies. This being said, we could have decided to apply the RF model to simulate groundwater elevation which can easily be converted to depth by subtracting it from the surface elevation. In general, we agree with the point raised by the reviewer that the groundwater elevation is more homogenous than its depth. This will be especially the case for shallow sandy aquifers, but in more complex geological settings, such as glacial tills, this is not necessarily the case. Here, a secondary shallow water table often follows the surface elevation. Simulating groundwater elevation instead of depth would be largely driven by the surface elevation and the complex influence of soil and topographical features may diminish. In order to test this assumption, we trained a RF model to simulate groundwater elevation, which we then converted to groundwater depth. The oob prediction is used to assess the accuracy and can be compared to the original RF model from the manuscript that directly simulates the depth. The results are presented in figure 1. The RF model trained against groundwater elevation shows more scatter/deviation, which is also underlined by the statistics. The accuracy of the oob prediction of the original RF model was: R^2=0.56, RMSE=1.13 m and MAE=0.76 m. The groundwater elevation RF model was: R^2=0.43, RMSE=1.28 m and MAE=0.80 m. The

scores of all three metrics worsened, which strengthens our original decision to simulate the groundwater depth. This comparison will be mentioned in the revised discussion section.

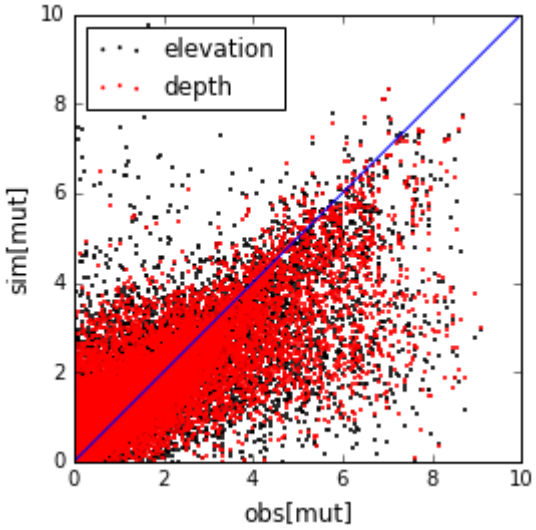

Figure 1 The oob prediction of the original RF from the manuscript, trained against groundwater depth, is compared with an alternative RF model, trained against groundwater elevation. The oob prediction of the latter was converted groundwater depth.

Secondly, I would like to comment on the use of a sine function to model an annual minimum event. I think it is a useful and generally robust way to address the issue of working with limited data. However, the method could be improved upon in a number of ways. Firstly, the maximum of the curve does not match the maximal observed water levels. The authors could therefore have calculated the uncertainty related to the sine model and, ideally, used these uncertainties in the Random Forest model. The authors already state this in the manuscript, but my second comment is related to the same issue. For training locations with sparse data, the authors used the maximum of the sine curve, but for training locations with more observations, the authors used observed maximum water levels. This choice muddles the results, both in terms of the predicted values and their accuracies. Is it a map of the expected minimum depth to the shallow water table, averaged over a number of years? Or is it a map of an extreme event, observed only in some years? The mixture of training data makes this question difficult to answer.

Reply: The sinusoidal correction model was a necessary step to be able to fully capitalize all available observations of the shallow groundwater system. The training dataset comprises 14916 wells of which 392 have a long timeseries (more than 5 observations). In other words, 97% of the training data underwent a correction with the defined sine models. The muddling of the results that the reviewer refers to is therefore not that severe as 97% of the data are processed in the same way and should represent a minimum event that is expected to occur every year. The minimum depth at the remaining 3% may slightly disagree, but we excluded very dry years from our analysis (1992-1997 and 2018) that could potentially lead to large disagreements. For the two examples in Figure 2 from the manuscript, the observed minimum values match quite well the sine model, with a deviation of approximately 10-20 cm. This falls within the uncertainty of the measurement itself. Therefore, we believe that the applied sine correction is a robust and appropriate approach. We acknowledge the uncertainties related to this processing step. The variability at the 392 wells with long timeseries has been investigated to infer the variability (amplitude) of the sine models for 27 different hydrogeological units. Expert knowledge was supplemented for groups that were poorly informed; i.e. only few available wells. The defined amplitudes are uncertain as the 27 groups may represent a crude classification and some groups are only represented by a limited number of wells. As we already outline in the paper, a more physically-based correction is needed for future applications. Using a hydrological model

for the correction would allow us to differentiate between inter-annual variations of the groundwater amplitude at grid level. Also, we could differentiate between dry and wet years.
We carried out a test to address the reviewers comments. We trained a RF model against the 97% of the observations that underwent the sine correction and withheld the 3% with more than 5 measurements to test the model. We then compared these results with the oob prediction of the original model. As indicated by the figure below, the predictions match quite well with a few exceptions.

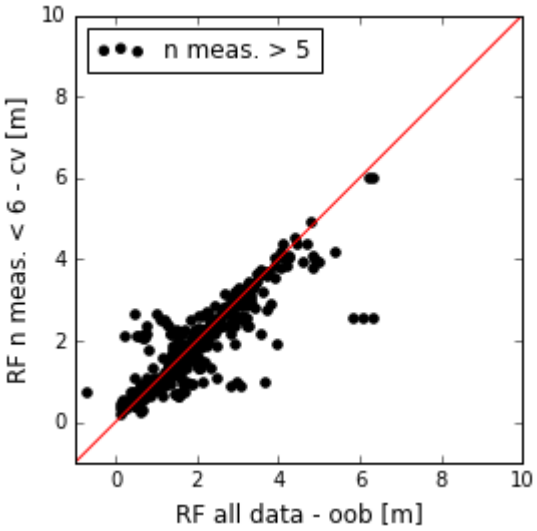

Figure 2 The plot shows predictions only for the 392 wells with long timeseries (more than 5 observations). The x axis presents the oob prediction of the original RF model from the manuscript. On the y axis, predictions of the 392 wells based on a RF model trained against the remaining 97% of the data with less than 6 observations.

Below, the same data is plotted against the observations. The two RF models differ, but the fact that no systematic difference is present strengthens our argument to use both, sine corrected observations and true observations, in the modeling process.

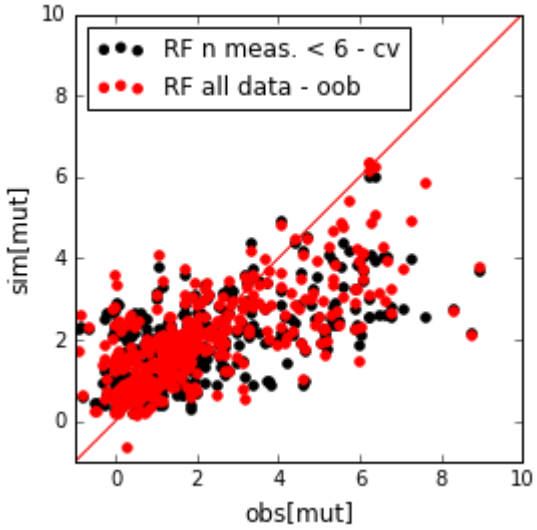

Figure 3 Same data as Figure 2, but plotted against the observations.

Thirdly, I have concerns about the way that the authors assess the accuracy of the predictions. The training dataset shows a high degree of clustering. Therefore, when the authors use the out-of-bag predictions for assessing the accuracy, the points used for

assessing the accuracy will be located close to the training points used for making the predictions. It is very likely that the values are spatially autocorrelated, and the stated accuracy is therefore probably not representative for the study area as a whole. I would expect the accuracy to be lower for the parts of the study area that do not have a high density of observations. A spatially structured accuracy assessment, as proposed for example by Muscarella et al. (2014), would most likely provide a more representative accuracy assessment. Furthermore, I am wondering if the authors used all the training points for the predictions. The training dataset contained both groundwater and surface water observations. However, the aim is not to predict surface water levels, and I would therefore say that one could justify removing the surface water points from the out-of-bag predictions when assessing the accuracy.

Reply: Inspired by this comment we prepared a 10-fold cross validation test to supplement the reported oob accuracy. Here we applied a standard method to partition the data randomly without considering the locations of the samples, as proposed by Muscarella et al. (2014). We agree that some of our data is clustered around cities or infrastructure projects. However, given a spatial autocorrelation of the RF residuals of 200 m (see variogram figure in the manuscript), we believe that with a simulation resolution of 50 m, we do not have to consider the spatial autocorrelation of the data for the partitioning for the 10-fold validation test.

We agree with the reviewer that the additional surface water observations with a groundwater depth of 0 m should not be included in the accuracy assessment. This was already considered in the submitted mansucript and will be stated more clearly in the revised version.

For the 10-fold cross validation test, 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 4. Figure 5 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 compared 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

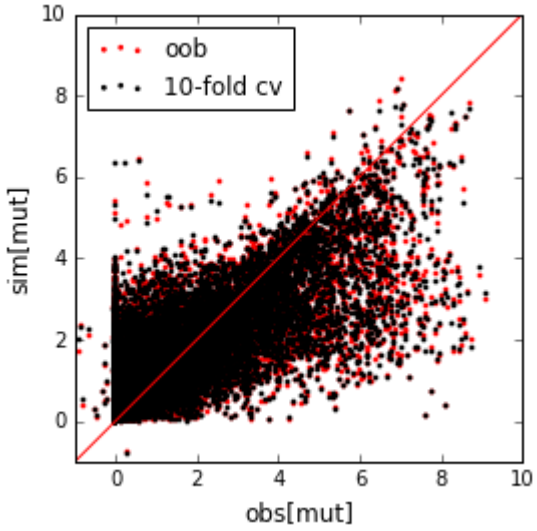

Figure 4 Comparison of oob prediction and 10-fold cross validation against observations.

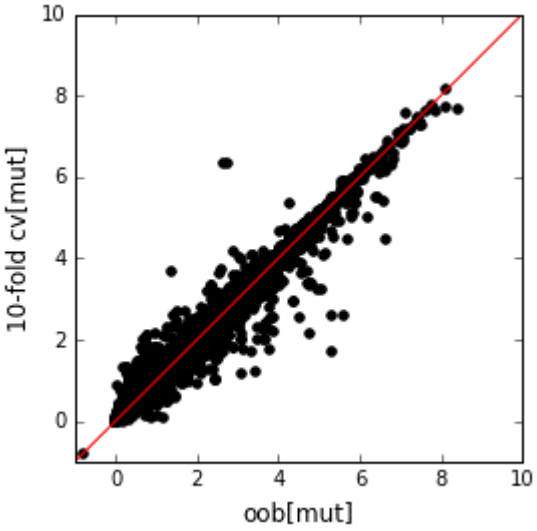

Figure 5 Direct comparison of oob prediction and 10-fold cross validation.

Fourthly, I very much like the way that the authors handle covariate importance. Being able to assess covariate importance in geographic space is potentially extremely useful, for both researchers and end users. However, I do not think that decrease in R2 is the best measure of covariate importance. One can potentially obtain a high R2, even if the absolute values are inaccurate. A better choice would therefore be to assess the relative change in a measure that accounts for absolute values, such as RMSE, Lin's concordance or the Nash-Sutcliffe efficiency.

Reply: We appreciate the reviewer's thoughts on our spatial covariate importance analysis. We need to clarify that the analysis on the prediction dataset which allows us to map covariate importance in space, uses the absolute difference between the permuted prediction and the original prediction. In this way, the reviewer's concerns were already considered in the original submission. For the purpose of the review we tested if the squred differences could give another result of the spatial covariate importance. This was not the case and the differences to the original results were minor. Therefor we decided not to include the results here or in the revised manuscript.
The decrease in R^2 was used to quantify covariate importance for the training dataset. Here we completely agree with the reviewer that the applied metric may not be the most suitable one. Therefore we tested the

analysis with the RMSE instead and results are shown in the figures below. The two figures below show the results based on the original assessment of the R^2 and the newly tests 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 a sensitivity analysis based on RMSE was made and that it yielded the same conclusions in terms of covariate sensitivity ranking.

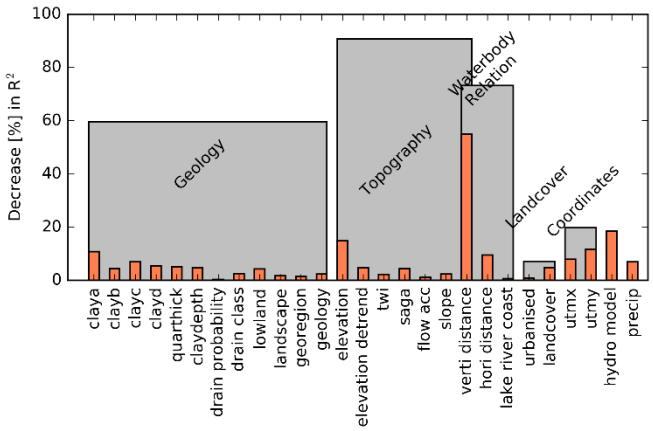

Figure 6 Covariate importance based on R2. Same as Figure 5 in the manuscript.

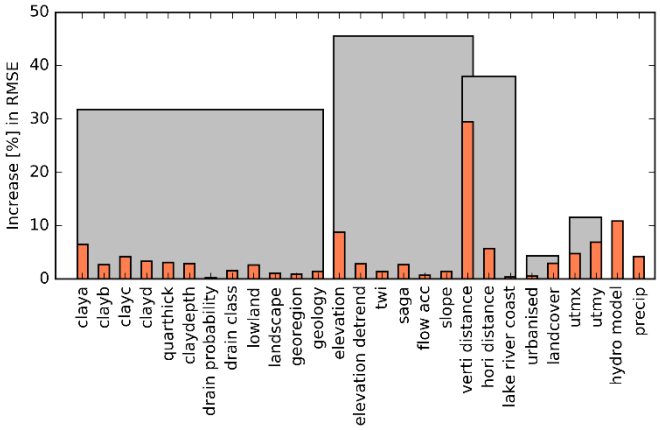

Figure 7 Covariate importance based on RMSE

Fifthly, while I appreciate that the authors assessed the uncertainties of the predictions in two different ways, I do not think that combining them is justified. The theoretical basis for the approach seems scarce. Both the RF uncertainties and the residuals used for kriging relate to the same model, and it is therefore a stretch to say that they are independent. Furthermore, quantile regression forest should be able to assess uncertainties quite accurately without any further elaboration, as shown for example by Rudiyanto et al. (2018). I think a large part of the spatial autocorrelation in the residuals would disappear, if one takes into account the uncertainties related to the RF predictions. The uncertainties in the predictions make the residuals uncertain as well, which complicates regression-kriging. When experimenting with techniques, as the authors do, it is important to set aside an independent part of the dataset to be able to assess the accuracy of the estimated uncertainties. However, the authors do not do this, and it is therefore impossible to assess if the error propagation actually leads to a better estimate of the uncertainties. Unless the authors can adequately adress these shortcomings, the section on error propagation should be removed. I am also

wondering why the authors used the out-of-bag residuals and not the residuals from the actual RF predictions. I have not seen any other studies using out-of-bag residuals for regression-kriging, and the authors should elaborate on their reasons for this choice.

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.

It is correct that we have used the oob predictions for the residual kriging and that this does not seem to be common practice in the literature. The RF model we apply is fully expanded until each leaf contains only a single data point. In this way, the standard RF prediction will be very close to the actual observation, as the observation value will be included in ~63% of the trees. This would lead to much lower residuals that do not really represent the generalization error. In this way, it would make no sense to interpolate the residuals for the standard RF prediction to unsampled locations. This assumption will be mentioned in the method section of the revised manuscript.

Sixthly, the authors use the hydrological DK-model as a covariate in the random forest model. I am wondering if the training points used in the RF model were also used for calibrating the DK-model. If this is the case, it creates a risk of circular logic, as the covariate contains information on the target variable at the location of the training points.

Reply: This is correct, some of the data was also used for the calibration of the DK model. However, for the purpose of this study, we have collected additional data from various sources (municipalities, region and consultancies) that were not yet in the national database and thus not been used for the calibration of the DK model. Further the shallow observations have so far received a low weight in the calibration of the DK model. Therefore, we do not believe that this problem will be a limitation of using the DK model as covariate in our model.

Seventh, the authors state that the sine model used to estimate extreme events could be replaced by an updated version of the DK-model. While I agree that this would improve the estimate of extreme events, it would also introduce another potential source of circular logic, if the DK-model was still used as a covariate. The approach would therefore need to be implemented with great care in order to avoid this.

Reply: We still believe that a more physically based correction of the observations is the way to move forward. Nevertheless, we agree with the reviewer that this can potentially lead to some issues of circular logic. We decided to remove this outlook from the manuscript as this is not really related to any of the results presented and thorough testing would be required before such a method could be implemented.

Lastly, I would like to comment on the use of the term "validation" for accuracy assessment. This is a general concern with the literature as much as a comment on this manuscript in particular. Oreskes (1998) argues that a quantitative model of a complex natural system cannot be considered as truly "validated" until it is used. For example, a conceptually flawed model can still provide good accuracies. The issue becomes even more relevant for machine learning models, where the parameters represent only patterns in the data, not physical processes. Strictly speaking, a machine-learning model can therefore never be truly valid, although it may be accurate and useful. To emphasize this point, I will mention Fourcade et al. (2018), who accurately mapped species distributions with entirely nonsensical covariates. I will encourage the authors to consider these points when discussing the accuracy of the predictions.

Reply: We are very much aware of the discussion on the term validation initiated by Oreskes (1998). In the revised manuscript we will downtown the term validation and use terms like "accuracy assessment" or "model evaluation" instead. However, we believe it is beyond the scope of this paper to discuss if machine learning models can be considered valid after being successfully evaluated against independent observations.

Technical corrections and stylistic suggestions

Reply: We appreciate this thorough technical/editorial review. We agree to all points raised by A.B. Møller and will updated the revised manuscript accordingly.

Generally, the authors refer to "traditional physically-based modelling" several times in the manuscript. I think "conventional" would be more adequate than "traditional", as science has conventions, not traditions. Tradition is a cultural phenomenon. Indeed, in most cases both "conventional" and "traditional" are redundant, as "physically-based modelling" accurately describes what the authors refer to, without any further need of clarification.
Page 2:
L5: "There exists a broad relevancy of the shallow groundwater" –> "The shallow groundwater has a broad relevance"
L9 – L10: "energy partitioning" –> "energy balance"
L13: "The shallow groundwater is also of importance in the urban context" –> "The shallow groundwater is also important in urban contexts"
L19: "a 100 year event with respect to today's average conditions" –> "a 100-year event relative to present average conditions"
L21: "high permeable" –> "highly permeable"
L28: "which hinders to conduct thorough calibration, sensitivity and uncertainty analysis at high resolution" –> "which hinders thorough calibration, and sensitivity and uncertainty analyses at high resolution"
L29: "Further, there exists a general difficulty to parameterize subsurface processes regardless the scale" –> "Furthermore, it is difficult to parameterize subsurface processes regardless of the scale"
Page 3: L3: "Hydrology" –> "hydrology"
L16: "mode" –> "model"
L16: "Before machine learning techniques can build the toolbox of future's environmental decision making" –> "Before machine learning techniques can be considered as a toolbox for environmental decision making"
L25: "Opposed" –> "However"
L29: "or" –> "and"
Page 4:
L3: "The study area encompasses a large part of the Danish peninsular, which is located in Northern Europe (54.5–57.8_N and 8.0–10.9_E) and referred to as Jutland." –> "The study area encompasses a large part of the Jutland peninsula, located in Denmark in northern Europe (54.5–57.8_N; 8.0–10.9_E)."
L5: "the sequence" –> "a sequence"
L6 – L8: The clay contents in eastern Denmark are not very high (10 – 20% for the topsoil). They are higher than the clay contents in western Denmark, but not relative to other areas in the world. It would be more accurate to say that the texture is loamy or that the clay contents are moderately high.
L8: "Weichselian sandy outwash plains" –> "sandy Weichselian outwash plains"
Page 5:
L6: "well data [. . .] was" –> "well data [. . .] were"
Page 6:

L6: "coast" –> "the coastline". This should be the case throughout the manuscript. Also "coastline" ! "the coastline".
Page 8:
Table 1: Lowland classification and landscape typology should refer to Madsen et al. (1992).
Table 1: "Drain probability" –> "Probability of artificial drainage"; "Drain class" –> "Soil drainage class".
Page 9:
L13: Bootstrap samples on average contain 63.2% of the data, not 66%.
L25: "The concept of covariate permutation allows to assess the importance of each covariate" –> "Covariate permutation allows an assessment of the importance of each covariate"
Page 12:
L20: "visual" –> "visible"
Page 13:
L2 – L3: Delete "was evident".
Page 17:
L21: "clear a shortcoming" –> "a clear shortcoming"
Page 19:
L3: "that region" –> "the study area"
Page 20:
L14: "allows to model" –> "enables"