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. Indication of changes made to the revised manuscript in red 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.

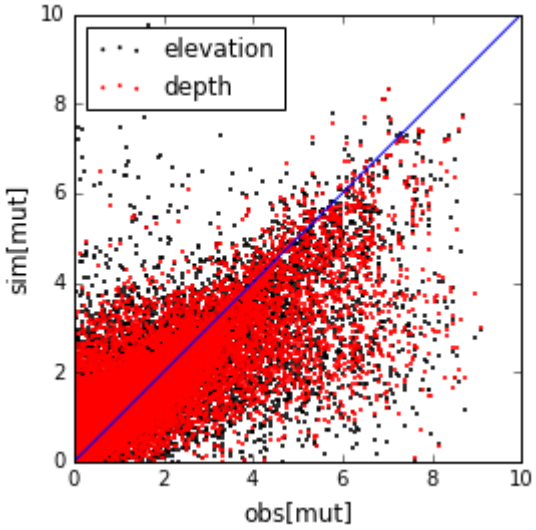

Figure 1 The oob prediction of the original RF from the manuscript, trained against groundwater depth, is compared with an alternative RF model, trained against groundwater elevation. The oob prediction of the latter was converted groundwater depth.

The revised manuscript contains a discussion on the abovementioned point on page 18, line 19 to page 19, line 4.

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.

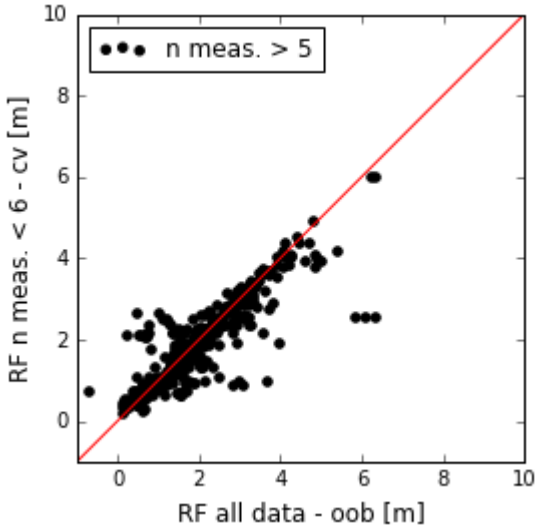

Figure 2 The plot shows predictions only for the 392 wells with long timeseries (more than 5 observations). The x axis presents the oob prediction of the original RF model from the manuscript. On the y axis, predictions of the 392 wells based on a RF model trained against the remaining 97% of the data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

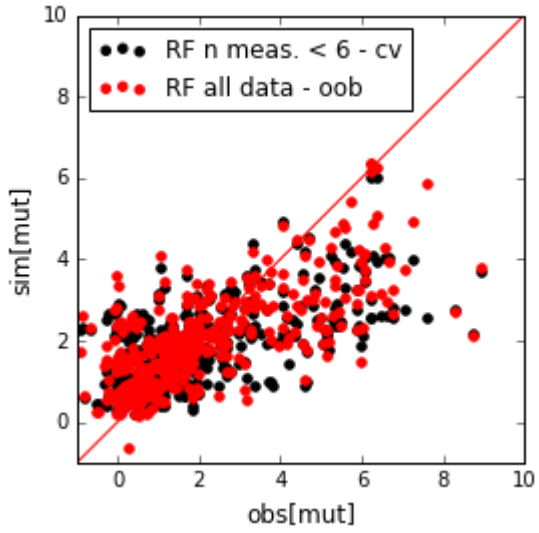

Figure 3 Same data as Figure 2, but plotted against the observations.

The revised manuscript contains an elaboration of the abovementioned point in the data section on page 6, lines 19 to lines 23.

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 700 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

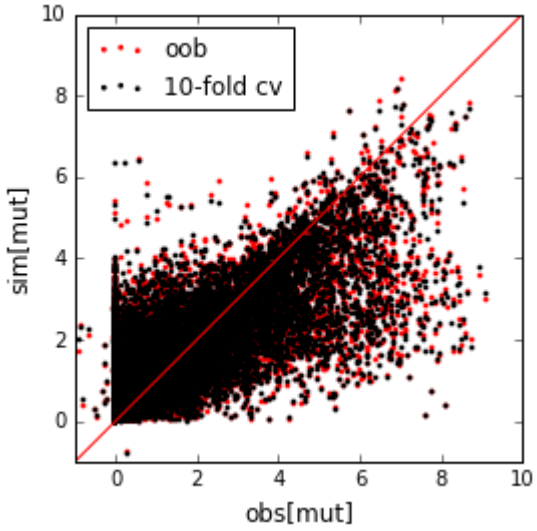

Figure 4 Comparison of oob prediction and 10-fold cross validation against observations.

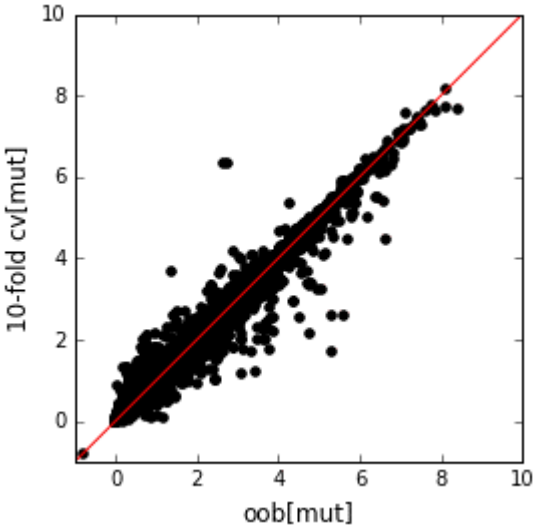

Figure 5 Direct comparison of oob prediction and 10-fold cross validation.

The revised manuscript contains a presentation of the cross validation test in the results section on page 12 line 15 to page 13, line 5.

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 squared 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.

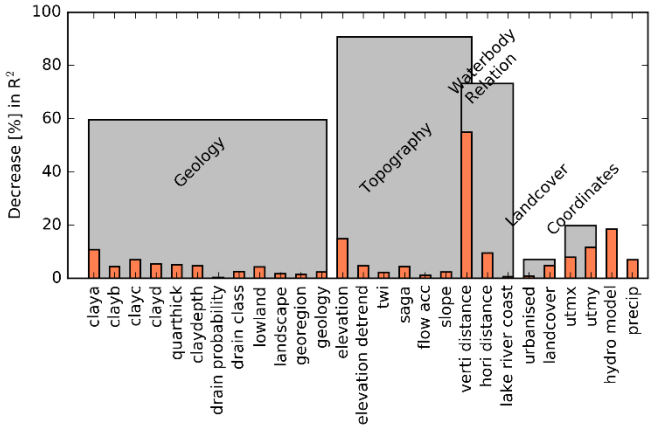

Figure 6 Covariate importance based on R2. Same as Figure 5 in the manuscript.

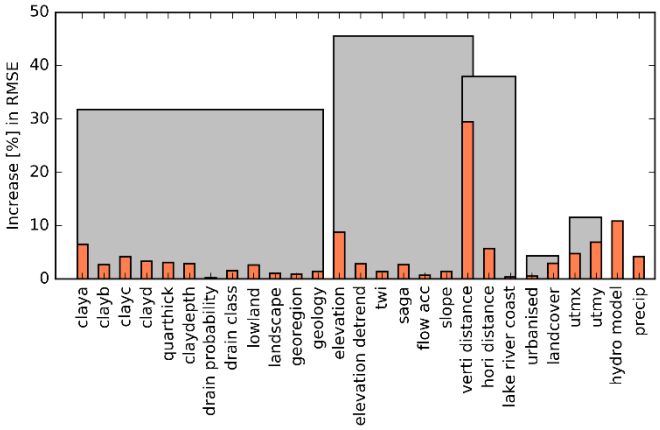

Figure 7 Covariate importance based on RMSE

The revised manuscript contains a discussion of the abovementioned point on page 15 line 3 to line 6. In addition, some clarification is added on page 14 line 6 to line 11, page 10 line 1 to line 2 and page 19 line 30 to line 33.

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.

The revised manuscript contains an updated method section on page 10, line 32 to page 11, line 2. Text related to the uncertainty propagation has been removed.

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.

No changes have been made.

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.
Text related to the above-mentioned point has been removed.

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.

Throughout the revised manuscript we have edited the term "validation".

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.

All suggestions have been incorporated.

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"

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.

The revised manuscript contains an updated data section on page 7, line 13 to line 15. The caption of Figure 1 has been changed as well.

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.

The revised manuscript contains a clarification of this point on page 7, line 21 to page 8, line 1.

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.

Section headings of 2.4 and 3.2 have been changed accordingly.

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.

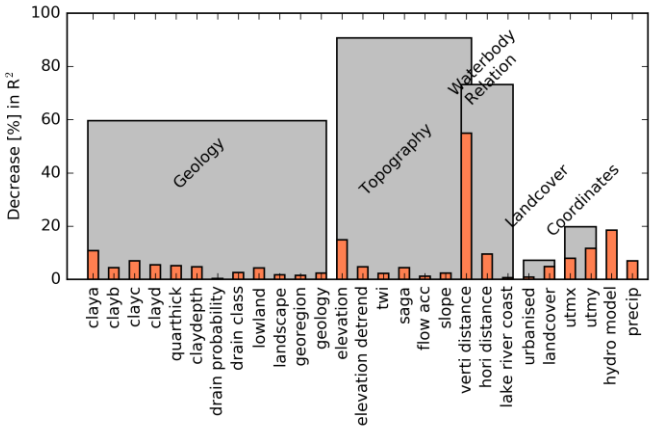

Figure 1 Covariate importance based on R2. Same as Figure 5 in the manuscript.

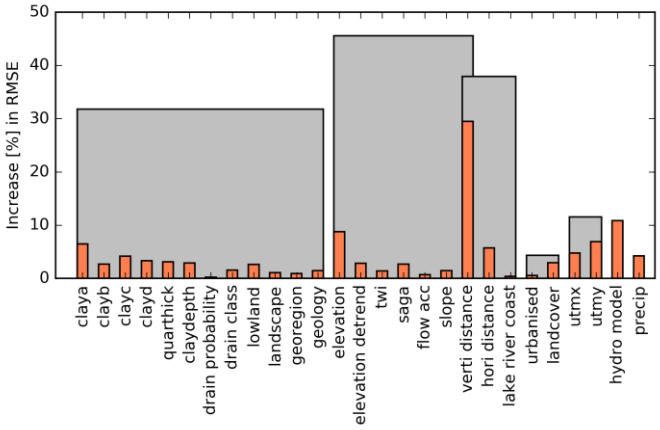

Figure 2 Covariate importance based on RMSE

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.

The revised manuscript contains a discussion of the abovementioned point on page 15 line 3 to line 6. In addition, some clarification is added on page 14 line 6 to line 11 and page 10 line 1 to line 2.

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.

The revised manuscript contains some additional clarification on this point on page 10, line 14.

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.

Clarification has been added at the end of section 2.6.

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.

Section 2.7 as well as related text in the results and discussion sections have been removed from the revised manuscript.

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.

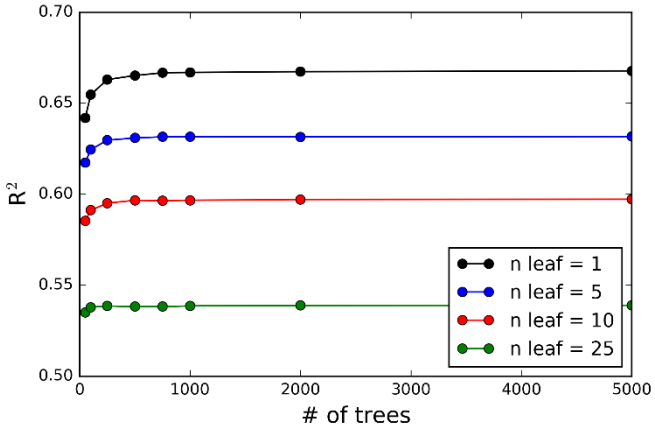

Figure 3 Initial testing of the RF hyper-parameters n_leaf and n_trees with respect to the R2 metric.

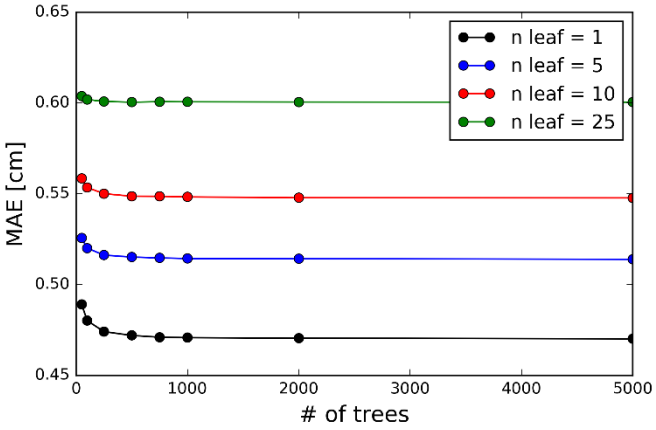

Figure 4 Initial testing of the RF hyper-parameters n_leaf and n_trees with respect to the MAE (mean absolute error) metric.

No changes have been made in the revised manuscript.

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

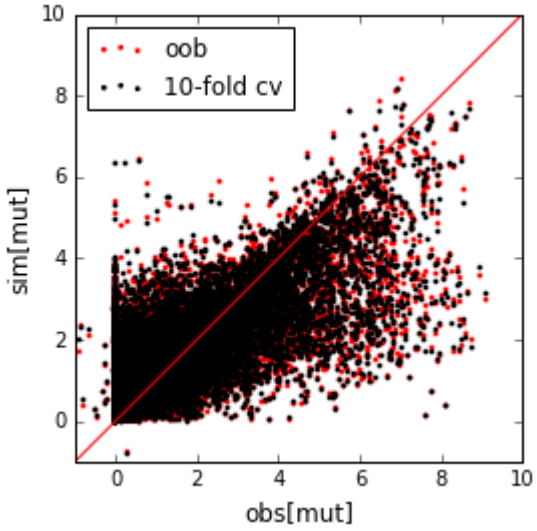

Figure 5 Comparison of oob prediction and 10-fold cross validation with respect to observations.

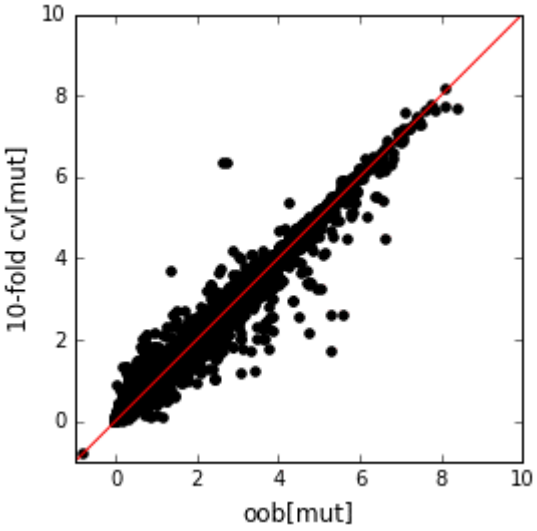

Figure 6 Comparison of oob prediction and 10-fold cross validation.

The revised manuscript contains a presentation of the cross validation test in the results section on page 12 line 15 to page 13, line 5. Also, more clarification has been added to the method section on page 9, line 24 to line 28.

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.

The Figure caption of Figure 3 has been changed as well as the results section on page 12 line 4.

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.

The revised manuscript contains a clarification on page 14 line 6 to line 11.

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 700 m. This limits the effect of the additional observations with a groundwater depth of 0 m to a close vicinity.

The revised manuscript contains a clarification on page 17 line 4 to line 8.

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 700 m (correlation length of the variogram) from an observations. Beyond 700 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.

The revised manuscript contains a clarification on page 17 line 9 to line 10.

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.

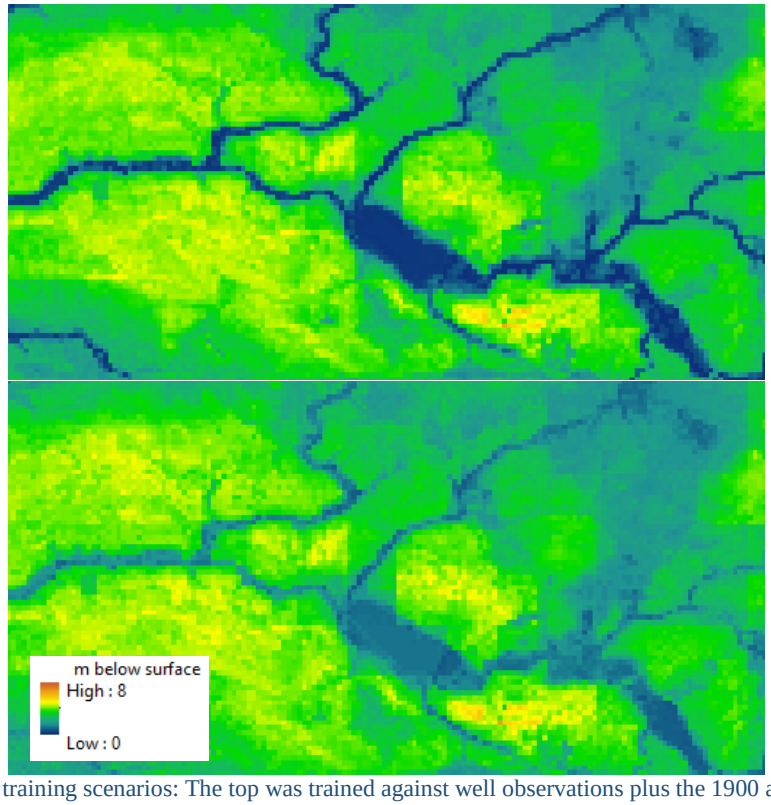

Figure 7 Results from two training scenarios: The top was trained against well observations plus the 1900 additional observation with a groundwater depth of 0 m. The bottom was trained exclusively against the well observations.

In the revised manuscript, the above-mentioned test is briefly mentioned on page 18, line 17 to line 18.

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.

The revised manuscript contains a discussion of the abovementioned point on page 15 line 3 to line 6. In addition, some clarification is added on page 14 line 6 to line 11 and page 19 line 30 to line 33.

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.

All suggestions have been incorporated.

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"

Modelling of the shallow water table at high spatial resolution using Random Forests.

Julian Koch[1], Helen Berger[2], Hans Jørgen Henriksen[1], Torben Obel Sonnenborg[1]

[1]Department of Hydrology, Geological Survey of Denmark and Greenland (GEUS), Copenhagen, 1350, Denmark
[2]COWI A/S, Lyngby, 2800, Denmark

Correspondence to: Julian Koch (juko@geus.dk)

Abstract. Machine learning provides a great potential to model hydrological variables at a spatial resolution beyond the capabilities of ~~traditional~~ physically-based modelling. This study features an application of Random Forests (RF) to model the depth to the shallow water table, for a wintertime minimum event, at 50 m resolution over a 15,000 km^2 large domain in Denmark. In Denmark, the shallow groundwater poses severe risks of groundwater induced flood events affecting both, urban and agricultural areas. The risk is especially critical in wintertime, when the shallow groundwater is close to terrain. In order to advance modelling capabilities of the shallow groundwater system and to provide estimates at scales required for decision making, this study introduces a simple method to unify RF and physically-based modelling. Results from the national water resources model in Denmark (DK-model) at 500 m resolution are employed as covariate in the RF model. Thereby, RF ensures physical consistency at coarse scale and fully exhausts high-resolution information from readily available environmental variables. The vertical distance to the nearest waterbody was rated the most important covariate in the trained RF model followed by the DK-model. The ~~validation~~ evaluation test of the trained RF model was very satisfying with a mean absolute error of 7~~9~~6 cm and a coefficient of determination of 0.5~~5~~6. The resulting map underlines the severity of groundwater flooding risk in Denmark, as the average depth to the shallow groundwater is 1.9 m and approximately 29 % of the area is characterised with a depth less than 1 m during a typical wintertime minimum event. This study brings forward a novel method to assess the spatial patterns of covariate importance of the RF predictions which contributes to an increased interpretability of the RF model. Quantifying uncertainty of RF models is still rare for hydrological applications. Two approaches, namely Random Forests Regression Kriging (RFRK) and Quantile Regression Forests (QRF) were tested to estimate uncertainties related to the predicted groundwater levels. ~~This study argues that the uncertainty sources captured by RFRK and QRF can be considered independent and hence, they can be combined to a total variance through simple uncertainty propagation.~~

1 Introduction

The shallow groundwater, defined as the uppermost water table, is a key state variable of the hydrological cycle having a wide range of vital implications on human health, terrestrial ecosystems, food security and energy production (Gleeson et al., 2016) . Following Fan et al. (2013), up to one third of the global land area is affected by the shallow groundwater, being either

directly groundwater-fed or having the water table or capillary fringe within plant rooting depths. In many regions of the world, groundwater aquifers are being depleted extensively by unsustainable anthropogenic activities (Richey et al., 2015). In addition, climate change affects groundwater recharge and storage which, in many cases exacerbate the resilience of shallow groundwater systems (Ferguson and Maxwell, 2010; Rodell et al., 2018).

~~There exists a broad relevancy of t~~The shallow groundwater has a broad relevance which expands beyond the hydrological science. For instance, Kahlown et al. (2005) and Zipper et al. (2015) studied the dependency between crop yield and the water table. They concluded that, in many agricultural settings, the groundwater played an essential role in meeting the crop water requirements. However, water tables too close to the surface resulted in reduced yields and both studies identified an optimal water table between 1 and 2 m below surface. Moreover, several studies highlighted the controlling mechanisms that the water table has on the energy ~~partitioning~~ balance at the land surface inferring a link to the latent heat flux and the delineation of water- and energy-limited ecosystems (Kollet and Maxwell, 2008; Maxwell and Condon, 2016). Other studies stressed the connections between groundwater and the near surface climate through coupled numerical modelling experiments (Larsen et al., 2016; Wang et al., 2018). The shallow groundwater is also of importance in ~~the~~ urban contexts (Bricker et al., 2017), with special focus on urban flooding which can be directly induced or indirectly intensified by high groundwater levels (Jankowfsky et al., 2014; Kreibich and Thieken, 2008; MacDonald et al., 2012). Moreover, MacDonald et al. (2010) and Upton and Jackson (2011) have studied the underlying processes, estimated return periods and mapped risk of groundwater flooding events.

In Denmark, the quantitative status of shallow groundwater systems is challenged by climate change and groundwater abstraction (Henriksen et al., 2008; Karlsson et al., 2016). In more detail, Kidmose et al. (2013) demonstrated that groundwater levels are expected to rise by up to 1.5 m for a 100 year event ~~with~~ relative to present average~~respect to today's average~~ conditions. Similar findings were presented by van Roosmalen et al. (2007) who quantified regional differences across Denmark in the projected change of groundwater levels depending on soil types with more profound increases in highly permeable sandy soils. Moreover, Henriksen et al. (2012) analysed climate change effects on the shallow water table over Denmark for mean and max conditions for nine different climate models and identified changes of at least 0.5 m for 26 % of Denmark. This finding represented the median change across the nine applied climate models.

The abovementioned problems call for comprehensive modelling tools that can support environmental decision making aiming at tackling today's and future's challenges related to the shallow groundwater. Spatial scales which are relevant for society and required for adequate decision making can typically not be provided by numerical, physically-based models alone. This limitation is mainly related to the fact that such models are computationally very expensive which hinders ~~to~~thorough conduct ~~thorough~~ calibration, sensitivity and uncertainty analysis at high resolution (Asher et al., 2015; Stisen et al., 2017). Further~~more~~, it is difficult to ~~there exists a general difficulty to parameterize~~parameterize subsurface processes regardless the scale (Beven et al., 2015). Moreover, the wealth and detail of hydrological data is under continual growth (Chaney et al., 2018) and the resulting big data is often not harnessed optimally in existing modelling frameworks (Best et al., 2015; Nearing et al., 2016). As outlined by Reichstein et al. (2019), machine learning will play an essential role in advancing ~~traditional~~ current modelling systems by integrating machine learning and numerical models. The development and testing of such hybrid models,

complementing benefits of physically-based models and machine learning, has gradually gained more attention in recent years in the hydrological community. A roadmap toward machine learning facilitated discoveries of hydrological systems has been outlined by Shen at al. (2018) and will likely play an eminent role in the future of ~~H~~hydrology. More generally, Karpatne et al. (2017) coined the paradigm "theory-guided data science" which comprises a diverse list of approaches with which physically-based models and machine learning can be combined. All three abovementioned references focus on the coupling of physically-based models with the versatility of data driven modelling frameworks. In more detail, they identified that the interpretability of machine learning models is among the main challenges for the successful adoption of big data technologies in the hydrological science.

This study highlights the applicability of machine learning, namely, the Random Forests (RF) algorithm (Breiman, 2001), to model the depth to the shallow groundwater at regional scale at high spatial resolution. The aim is to produce a map that captures an extreme wintertime condition representing a minimum depth to the water table. Such an event can potentially induce groundwater flooding which poses risks related to infrastructure and agriculture. Thus, the resulting high resolution map will be a versatile screening resource for environmental decision making and climate change adaption planning. The proposed RF model draws on the Danish national water resources model (Højberg et al., 2013) which provides a coarse estimation of the depth to the shallow water table. In this way, the RF model utilizes the coarse prediction of the physically-based model to ensure overall physical consistency, which may not be granted by the RF model alone. Hence, this study tests a simple hybrid model integrating the output of a numerical model in a machine learning framework. Before machine learning techniques can be considered as a standard tool for~~build the toolbox of future's~~ environmental decision making and planning, methods to conduct comprehensive sensitivity analysis and uncertainty assessment need to be developed and tested thoroughly. In order to advance this field of hydrological research, this study compares two different methods to quantifying uncertainty of a RF model, namely Random Forest Regression Kriging (Hengl et al., 2015) and Quantile Regression Forests (Meinshausen, 2006). Furthermore, this study features a novel methodology to quantify covariate importance, and thereby the sensitivity of the model inputs, which ultimately helps to better comprehend and interpret the RF prediction.

Numerous studies have already successfully employed machine learning techniques to predict the temporal dynamics of the groundwater system based on artificial neural networks (Banerjee et al., 2009; Daliakopoulos et al., 2005; Shiri et al., 2013; Yoon et al., 2011) or other machine learning ~~techniques~~methods (Fallah-Mehdipour et al., 2013). ~~Opposed~~However, opportunities to utilize data driven modelling to assess the spatial dimension of the water table have so far not been fully exhausted. Fienen et al. (2013) are among the few that utilized a machine learning technique to map the depth to the groundwater in space (i.e. Bayesian network). Machine learning has already been applied to model other groundwater related variables such as, nitrate concentration (Nolan et al., 2015; Tesoriero et al., 2015), arsenic concentrations (Erickson et al., 2018; Winkel et al., 2011) ~~or~~ and redox conditions in the subsurface (Close et al., 2016; Koch et al., 2019), and the potential to map the depth to the water table is tangible.

The four main objectives of this paper are as follows: (1) to train a RF model that is capable to predict the depth to the shallow water table at high spatial resolution, (2) to outline a simple and generic method that unifies a physically-based model and

machine learning, (3) to conduct a comprehensive sensitivity analysis to better interpret the RF model prediction, (4) to assess the uncertainty related to the RF model based on two different approaches.

2 Methods

2.1 Study Area

The study area encompasses a large part of the ~~Danish~~ Jutland peninsular, which is located in Denmark ~~in N~~northern Europe (54.5–57.8°N and 8.0–10.9°E) ~~and referred to as Jutland~~. The extent of the study area amounts to approximately 15,000 km^2 and its general surficial geological setting is illustrated in Figure 1. The landscape of Jutland was formed by ~~the~~ a sequence of Pleistocene glaciations and postglacial processes. The geology of the eastern part is dominated by Weichselian moraine sediments with ~~high~~ moderate clay content, whereas the west is characterized by moraine sediments originating from the Saalian age (Hill Island) intertwined by sandy Weichselian ~~sandy~~ outwash plains.

2.2 Data

This study aims at modelling the depth to the shallow water table at 50 m spatial resolution using a machine learning modelling framework~~.~~ Disregarding the prevailing temporal variability of groundwater dynamics close to the surface, we chose to model an extreme event that characterises a minimum depth, expected to arise every year. Based on the climate in Denmark, such an event normally occurs towards the end of winter, when shallow aquifer systems are replenished after several months of typically high rainfall and low evapotranspiration. Applying machine learning to model an extreme event of a highly dynamic variable poses distinct challenges to the training dataset. Long timeseries of groundwater head measurements are scarce, and shallow groundwater time series ~~is~~ are even more rare. In fact, many shallow wells, with screens within the uppermost 10 meters, provide just one to a few observations in total. In order to capitalize these low frequency sampled wells, we have developed a method to transform any given observations to an expected high water table. For this transformation, sinus curves were defined with varying amplitudes capturing the annual dynamics of the shallow groundwater for various hydrogeological settings. The workflow is described in more detail below.

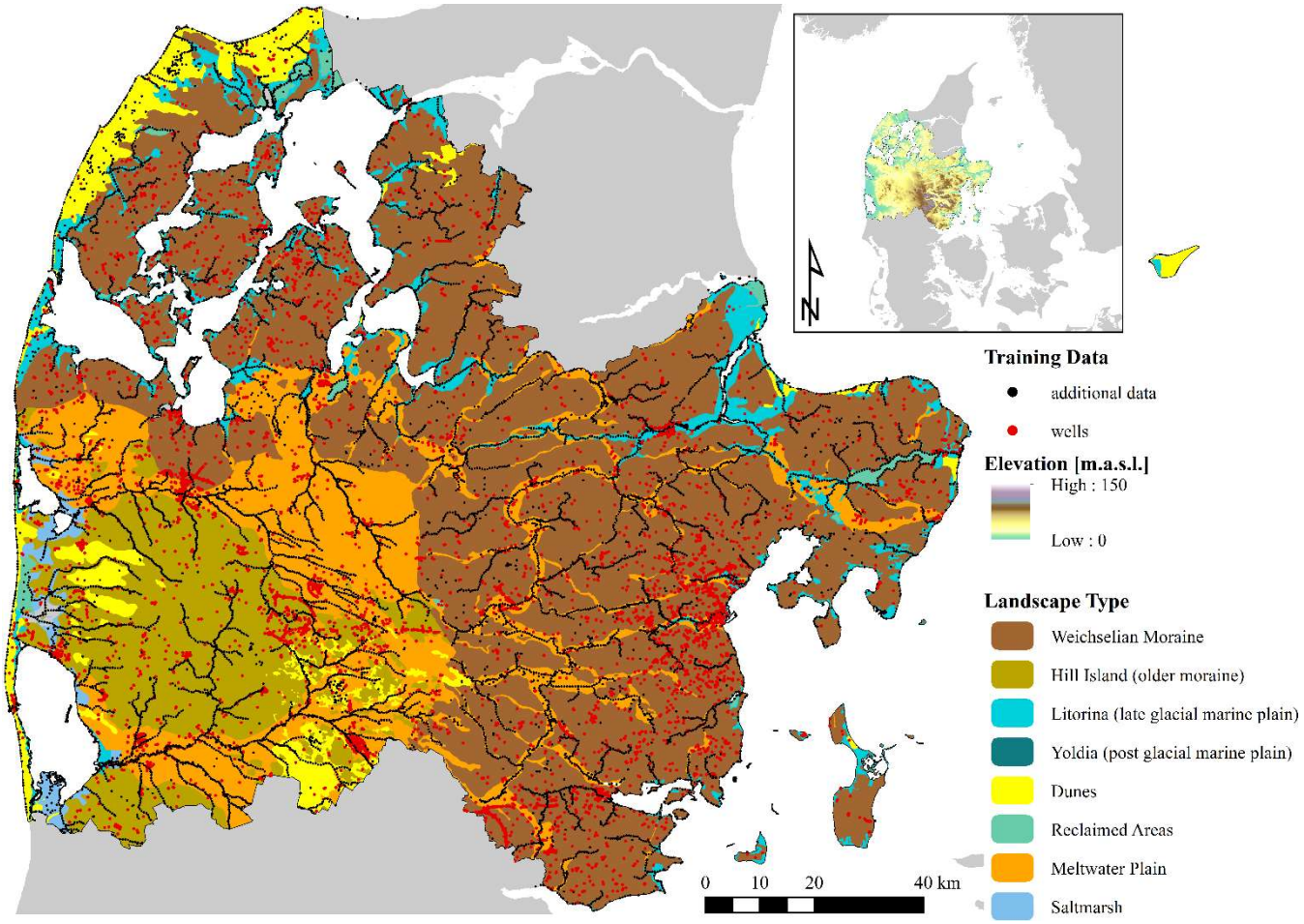

Figure 1: The study site is located in central Denmark. The overview figure, in the uppermost right corner, depicts the digital elevation model. The main map shows the predominant geological landscape types. The training dataset contains observations at ~15,000 wells. Additionally, ~~and~~ ~1~~5~~6,000 ~~additional~~ artificial ~~data points~~observations are placed along major rivers, lakes and the coastline are added to the analysis. The depth to the shallow groundwater is set to zero for the additional data.

First, well data, covering the entire model domain, ~~was~~ were extracted from the national database, Jupiter (Hansen and Pjetursson, 2011). Groundwater head observations from wells with a maximum filter depth of 10 m were assorted for a 20 year period between 1998 and 2017. Several constraints were applied to this initial extraction: (1) the mean water level may not be below filter depth, (2) the water levels may not exceed 5 m above surface, (3) the standard deviation of head observations may not be greater than 3 m and (4) the well may not be in operation. By applying these four constraints, 14,916 wells with one or more head observations were selected which approximately corresponds to a density of one well per km^2. Figure 1 shows the location of the wells.

Second, wells with more than five observations, of which 392 were present, were grouped according to their hydrogeological setting. Subsequently, their standard deviation was studied in more detail in order to define the sinus curve amplitudes for each of the groups. In total, 27 combinations, describing the general hydrogeological setting of a well, were assessed. These groups

were based on three categories with three sub-categories each, (1) permeability (high, low or unknown), (2) aquifer condition (confined, unconfined or unknown) and (3) proximity (near the coastline, near stream or other). The amplitude of the sinus curve was set to the 99% confidence interval and, under the assumption of normality, calculated as 2.576 times the standard deviation. Based on the analysis of the variability at 392 wells with long timeseries the average annual amplitude of the sinus curves varied between 0.5 m and 1.5 m depending on the hydrogeological setting. The largest amplitude was associated to wells with filters in low permeable sediment, under unconfined conditions and not in the vicinity of the coastline ~~coast~~ or streams. Low amplitudes were generally connected to wells closer than 100m to streams, lakes or the coastline~~coastline~~. The minimum and maximum of the sinus curves was set to arise in ~~the~~ mid February and mid August, respectively.

Third, the groundwater head value describing an extreme wintertime condition at each well was defined twofold. At wells with five or more observations the recorded minimum depth was used as input to the training dataset. Opposed, at wells with fewer observations, the predefined sinus model was applied to transform an observed minimum to an expected wintertime minimum. With this approach, the observed variability at wells with high quality data was used to infer a meaningful minimum value, describing an extreme wintertime situation, at wells with few observations. Figure 2 exemplifies this approach in more detail. The first two examples express wells with long timeseries, used to define the sinus amplitude corresponding to the well's hydrogeological setting. Here, the minimum observations were extracted as input to the training dataset. The bottom two examples depict cases with few observations where a sinus curve was applied to transform the observed minimum depth to an expected winter minimum condition. In cases where the transformation resulted in negative values, i.e. manifesting artesian conditions, the value was set to zero. This correction was considered meaningful as many of these wells were located in unconfined conditions. The resulting training dataset consisted of 14916 wells, of which 97% were corrected with sinus curves and at the remaining 3% the observed minimum depth was used. Figure 2 indicates that the minimum depths based on the measured values and sinus model may deviate for the 392 wells with long timeseries. However, the two examples in Figure 2 imply a deviation of 10-20 cm, which lies within the measurement uncertainty itself. This allows us to conclude that the proposed sinus model based correction is robust enough for our application.

There are two sources of uncertainty that were not considered in this analysis. First, the observational uncertainty related to the head values in the well database was not considered. Second, the sinus model used to traverse any given observation to an expected wintertime minimum neglecteds seasonal and inter-annual variability.

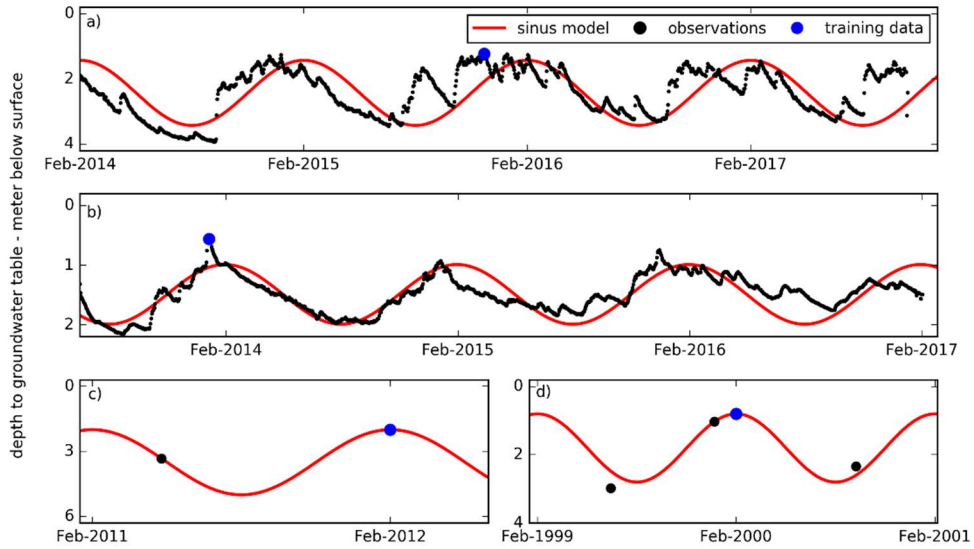

Figure 2: Four examples showing how the training dataset was derived. At wells with more than four observations, a) and b), the minimum daily observation was chosen. Examples a) and b) represent long timeseries and sinus curves with amplitudes of 1 m and 0.5 m, respectively, ~~that~~ were used to describe the annual variability. Examples c) and d) represent two cases with few observations. Here, sinus curves with predefined amplitudes, 1.5 m and 1 m, respectively, corresponding to the well's hydrogeological setting were applied to traverse the observed minimum depth to an expected wintertime minimum.

Additional observations were placed along streams, the coastline ~~coastline~~ and the centre points of lakes. At the given locations, the depth to the shallow groundwater was set to zero. This extension of the training dataset was found necessary in order to provide critical information to the machine learning model which was otherwise not conveyed in the borehole data alone. However, only a subset consisting of 1,900 random samples of the 16,210 additional observations was used for training of the machine learning model. This corresponds to approximately the same well density as found in the original training dataset, taking into consideration the combined area of streams, the coastline ~~coastline~~ and lakes in 50 m grid resolution. In this way, the information content of the well data and the additional data was balanced. Including all ~16k additional observations with a depth of zero in the training dataset would result in a biased model, because the average of the training dataset would not be representative for the study site. The complete dataset of additional observations was however utilized in the uncertainty analysis.

In Table 1, a list of the environmental covariates used to model the depth of the shallow water table is found. In total, 26 covariates were assembled as input to the machine learning model. This list comprises information on soil texture, drainage conditions, geology, topography based characteristics, waterbody proximity, precipitation, land cover, geographic location and outputs from a hydrological simulation with the Danish national water resources model (DK-model: Højberg et al., 2013). The native spatial resolution of the covariates varied, but all covariates were resampled to 50 m to be in agreement with the defined

output resolution. The waterbody proximity was expressed as both the vertical and horizontal distance to the nearest waterbody, which contained rivers, lakes and the coastline. The covariates were subdivided into six groups, i.e. geology, topography, waterbody relation, precipitation, land cover, coordinates and hydrological model. This subdivision was implemented in the sensitivity analysis of the machine learning model to eliminate correlations between covariates.

Table 1: Overview of the covariates used to model the shallow water table using RF.

Variable	Source	Group
Clay content; a layer: 0-30 cm	Adhikari et al., 2013	Geology
Clay content; b layer: 30-60 cm		
Clay content; c layer: 60-100 cm		
Clay content; d layer: 100-200 cm		
Quaternary thickness	Binzer and Stockmarr, 1994	
Depth to clay occurrence	Højberg et al., 2013	
~~Drain p~~Probability of artificial drainage	Møller et al., 2018	
Soil ~~D~~drainage class	Møller et al., 2017	
Lowland classification	Aarhus University – Danish Centre for Environment and Energy: The Danish SINKs project	
Landscape typology	~~(~~Breuning-Madsen and Jensen, 1992~~)~~	
Georegion classification	Adhikari et al., 2014	
Soil type	Geological Survey of Denmark and Greenland	
Digital elevation model	Danish Agency for Data Supply and Efficiency (SDFE)	Topography
Detrended digital elevation model		
Topographic wetness index	Böhner and Selige, 2002	
Saga wetness index		
Flow accumulation	SDFE	
Slope		
Vertical distance to nearest waterbody		Waterbody relation
Horizontal distance to nearest waterbody		
Waterbody (lake, river and the coastline~~coast~~) classification		
~~Precipitation~~	~~Danish Meteorological Institute~~	~~Precipitation~~
Degree of urbanization	Danish climate change adaption portal	Land cover
Land cover	CORINE Copernicus	
Coordinates (utmx)	N/A	Coordinates
Coordinates (utmy)		
DK-model; depth to max groundwater level	Henriksen et al. 2014~~(Højberg et al., 2013)~~	Hydro model
Precipitation	Danish Meteorological Institute	Precipitation

2.3 Random Forests

This study applied Random Forests (RF) regression to model the depth to the shallow water table at high spatial resolution at regional scale. RF was first proposed by Breiman (2001) and emerged to a prevalent modelling tool covering a wide range of geophysical and environmental contexts. These include, among others, digital soil mapping (Hengl et al., 2015; Ließ et al., 2012), estimating nitrate pollution in aquifers (Rodriguez-Galiano et al., 2014; Tesoriero et al., 2017), biomass estimation using satellite remote sensing (Mutanga et al., 2012), landslide susceptibility analysis (Youssef et al., 2016), mineral prospectivity mapping (Rodriguez-Galiano et al., 2015) or estimation of young water fractions across catchments (Lutz et al., 2018). RF has proven to provide high predictability for multivariate modelling of complex, non-linear variables and multiple benchmarking studies have documented the capabilities of RF to outperform other machine learning techniques (Nussbaum et al., 2018; Rodriguez-Galiano et al., 2015; Youssef et al., 2016). Like other data driven modelling approaches, training is an essential step in the RF model building process. Based on the training dataset, RF learns linkages between the covariates and the target variable at sampled locations, which then are generalized to make predictions at unsampled locations. The core of RF is an enhanced utilization of decision trees. More precisely, RF builds an ensemble of decision trees, where each tree recursively splits the training data into more homogenous groups. The formulation of the decision trees contains two elements of randomness with the aim to increase the diversity within the ensemble of decision trees. First, the concept of bagging is applied. Bagging is an ensemble technique, which generates a unique bootstrap sample of the original training dataset for each decision tree. Based on sampling with replacement, each bootstrap sample contains approximately ~~66~~ 63.2 % of the original training data. The average across the entire ensemble represents the final RF prediction. Second, only a subset of covariates is drawn upon when splitting the data during the process of decision tree building. This subset, usually 33 % of the available covariates, is selected randomly for each split. In combination, the two elements of randomness decrease the accuracy of a single tree; however the diversity between the trees increases which results in a robust prediction when averaging across all trees.

The bootstrapping procedure divides the training data into an in-bag part, which is used for building the decision trees, and an out-of-bag (oob) part, which is excluded from the training. This partitioning is unique for each tree of the ensemble and thereby provides a valuable internal cross validation test. In other terms, each tree can be ~~validated~~ evaluated with its own~~a different~~ oob sample and the average across all oob predictions allows to quantify the overall accuracy of the RF model. For the oob prediction, only samples retained from the training, thus out-of-bag, are considered when averaging. In this way, the entire ensemble of trees can be evaluated by applying the oob approach. In order to quantify the performance of the RF model we have applied the following metrics on the oob prediction: Coefficient of determination (R^2), mean absolute error (MAE) and root mean squared error (RMSE). We applied the python package Scikit-learn (Pedregosa et al., 2012) to conduct the RF modelling for this study.

2.4 Covariate Sensitivity ~~Analysis~~

The concept of covariate permutation allows to assess the importance of each covariate, acting as input to a RF model (Biau and Scornet, 2016). This can be understood as a sensitivity analysis which can help to better comprehend and interpret the trained RF model and to gain physical insights into the otherwise intransparent black box model. This is achieved by permuting each covariate at a time, while leaving the remaining covariates unchanged, and tracing the apparent decrease in oob ~~validation~~ evaluation metric~~,~~. ~~typically the coefficient of determination.~~ Typically, the coefficient of determination R^2 is used as metric, but other metrics could be consulted as well as the sensitivity may be metric dependent. This concept is common practice to assess covariate importance for a trained RF model (Ließ et al., 2012; Lutz et al., 2018). However, this analysis is limited to the training dataset and conclusions on which covariates dominate the prediction and how this varies spatially cannot be drawn. In order to gain insights into the spatial patterns of covariates importance, we have developed a novel method, which applies the abovementioned concept of covariate permutation on the prediction dataset instead of the training datatset. The aim of the sensitivity analysis is to identify a relative ranking of covariate importance for each simulation grid, which can ultimately provide an increased interpretability ~~for the users~~. The starting point of the analysis is the trained RF model and its prediction ~~for~~ at all simulation grids. Sequentially, each covariate is permuted, while leaving the remaining covariates unchanged, and the trained RF model is used to make a modified prediction. The difference between the modified and original prediction is recorded. The cycle of permutation and prediction is repeated n times until the mean difference across n permutations converges for each simulation grid. This is necessary, because a single permutation may allegedly result in no or minor change in covariate value at specific grids. Once the mean difference has converged, the covariates can be ranked with respect to their associated mean absolute difference for each simulation grid. ~~-~~ In order to map the spatial covariate sensitivity it is essential that the ranking is performed at each simulation grid. This ranking expresses the relative covariate importance and is the key result of the proposed sensitivity analysis. Maps showing the top ranks can be used to visualize the spatial patterns of the sensitivity of the RF model.

Typically, strong correlations are found between covariates, which may result in an alleged low importance when being permuted individually (Koch et al., 2019). In order to overcome this limitation we suggest a supplementary analysis that collectively permutes groups of covariates that are physically related.

2.5 Random Forests Regression Kriging

Extending RF with geostatistical methods is gaining popularity in the field of digital soil mapping (Guo et al., 2015; Hengl et al., 2015) and related environmental modelling studies (Ahmed et al., 2017; Li et al., 2011; Viscarra Rossel et al., 2014). Regression Kriging (RK) is a widely applied approach that combines a multiple linear regression (MLR) model with a geostatistical model of the MLR residuals (Hengl et al., 2007; Odeh et al., 1995). In order to integrate RF into RK, RF can simply replace the MLR model. In this way, RF provides an overall data-driven trend and the RF residuals can be interpolated using geostatistics. This results in a hybrid model that is commonly referred to as Random Forests Regression Kriging (RFRK).

To our knowledge, RFRK has not yet been applied with the purpose to predict a hydrological state variable such as groundwater head. RFRK can be expressed by

$$P_{RFRK}(s_0) = T_{RF}(s_0) + \hat{e}_{RF}(s_0), \quad \text{(eq.1)}$$

where $T_{RF}(s_0)$ is the RF prediction at location s_0 and $\hat{e}_{RF}(s_0)$ is the estimated residual at the same location. The sum of trend (T_{RF}) and residual ($\hat{e}_{RF}$) yields the final RFRK prediction (P_{RFRK}). This study utilizes kriging to interpolate the oob residuals of the RF model. We use the oob prediction instead of the overall RF prediction to compute the residuals, because the oob procedure provides a more realistic estimation of the generalization error. The overall RF prediction naturally exhibits a lower error than the oob predictions as the data was contained in the training. Thereby, the resulting error variance would be biased and can not be used to interpolate the error at unsampled locations. Kriging is a popular geostatistical technique for spatial interpolation that employs knowledge about the spatial autocorrelation of a variable, which can be captured by a variogram model. For the definition of a variogram model, the omnidirectional empirical semivariance (γ) is calculated by

$$\gamma(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} [e(s_i) - e(s_i + h)]^2, \quad \text{(eq.2)}$$

where *n(h)* marks the total number of data pairs at a given lag distance *h*. $e(s_i)$ represents the oob residual at location s_i and $e(s_i+h)$ is the residual separated by lag *h* from s_i (Matheron, 1963). A variogram model is fitted to γ to model the spatial autocorrelation structure of the oob residuals (Deutsch and Journel, 1998). The parameters defining a variogram model are type, range, sill and nugget. The R package Gstat (Pebesma, 2004) was applied for variogram modelling and kriging interpolation.

The addition of residual kriging to RF results in high accuracy at grids coinciding with observations. Furthermore, kriging quantifies the prediction uncertainty following the defined variogram model. Generally, the kriging variance is low in vicinity to data points and increases to the sill value once the distance to the nearest data point is beyond the range of the variogram model.

2.6 Quantile Regression Forests

Using ~~traditional~~ RF, the prediction is obtained by averaging across the ensemble of decisions trees. This disregards the distribution of the target variable originating from several hundreds to thousands of decision trees, which are typically necessary to build a robust RF model. Meinshausen (2006) developed the Quantile Regression Forests (QRF) method that analyses the quantiles of the distribution of the target variable at prediction grids. This results in an estimation of prediction uncertainty or prediction intervals. The latter is obtained by recording specific quantiles which mark the lower and upper confidence limits (Hengl et al., 2018). The adoption of QRF for hydrological variables is still gradually and only few studies documented its applicability (Francke et al., 2008; Zimmermann et al., 2014). To our knowledge QRF has not been applied to quantify uncertainty of groundwater level predictions. For this study, we utilized the RF functionalities from Scikit-learn (Pedregosa et al., 2012) to implement QRF.

~~**2.7 Error Propagation**~~

~~The use of error propagation allows combining several sources of uncertainty. In order to apply this concept, no significant covariance between the uncertainties may be present. Following this assumption, the uncertainties (σ) of different sources (x, y, etc.) can be combined by the following approach:~~

~~$\sigma_{Combined} = \sqrt{\sigma_x^2 + \sigma_y^2 + \cdots}$.~~ ~~(eq.3)~~

~~The key limitation is that the underlying assumption seldom holds (Refsgaard et al., 2007). Nevertheless, this propagation approach can facilitate a simple and powerful screening analysis (Kidmose et al., 2013).~~

3 Results

3.1 Random Forests Model

For the purpose of modelling the depth to the shallow water table at 50 m spatial resolution for an extreme wintertime minimum event, a RF model was trained using the 26 available covariates and groundwater head data. The training data comprised ~15,000 wells and 1,900 additional observations placed along streams, the coastline ~~coastlines~~ and in lakes. 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, trees were fully expanded and thus not pruned, the mean squared error was selected as criterion to identify the optimal data split and regression was chosen as method.

Figure 3 depicts the internal cross-validation test based on the oob samples of the well data. The oob prediction can be considered as an independent ~~validation~~ evaluation test and three performance metrics, i.e. coefficient of determination (R^2), mean absolute error (MAE) and root mean squared error (RMSE) indicated an overall good performance (Table 2). More than half of the variance contained in the training data was captured by the RF model, the MAE amounted to 7~~9~~6 cm and the RMSE was 1.13~~9~~ m. The density scatter plot in Figure 3 zooms into the top 6 m and it becomes apparent that very shallow observations (< 0.5 m) were systematically biased while deeper observations were estimated in good agreement, closely to the 1:1 line.

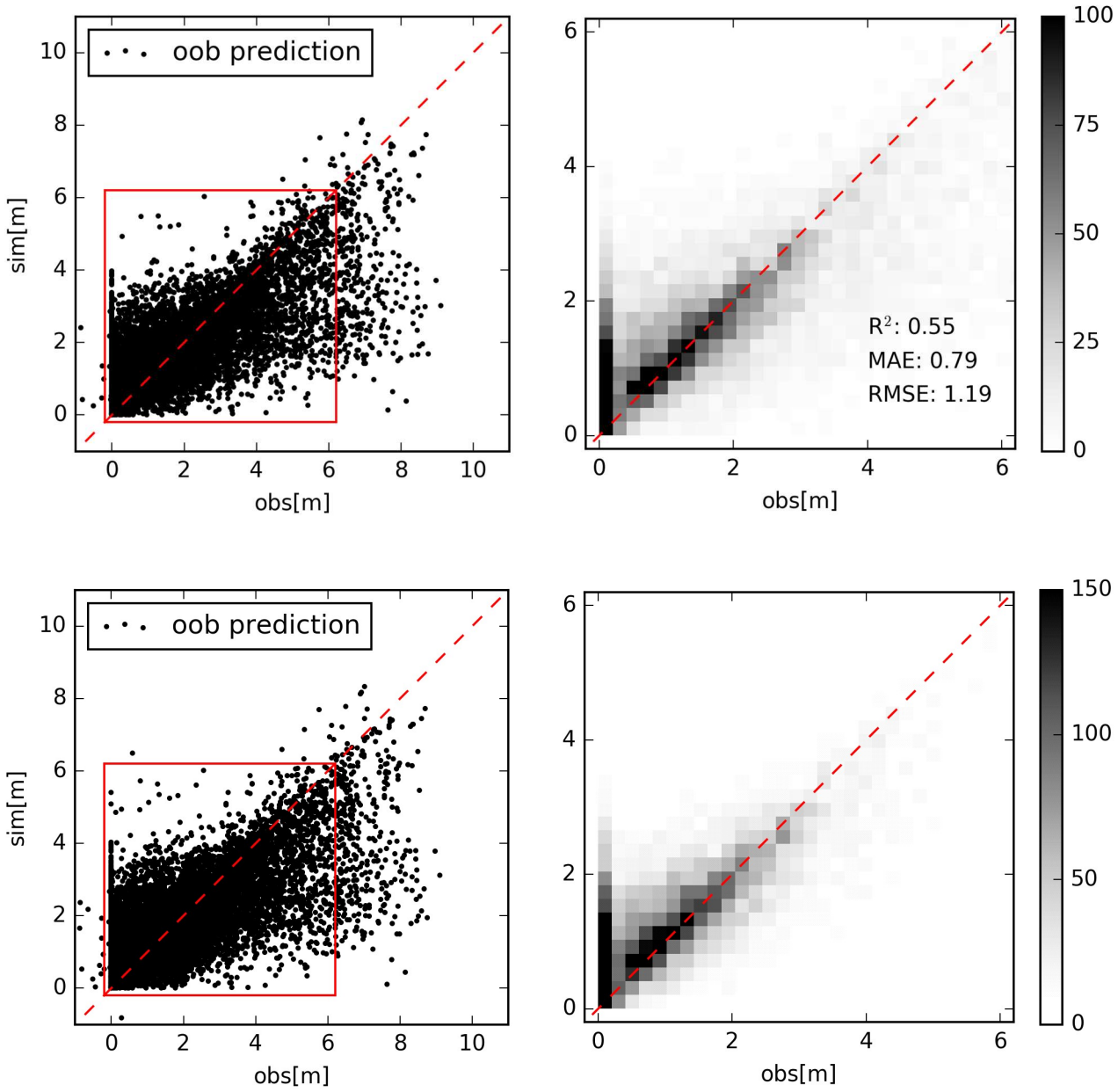

Figure 3: The RF ~~validation~~ accuracy assessment ~~test~~ was performed based on the out-of-bag sample technique. The axes depict the simulated and observed depth to the shallow water table. The left panel displays a standard scatter plot containing ~15~~7~~,000 ~~data~~ well data~~points~~. The 1,900 additional observations are excluded. The right panel shows a zoom (extent indicators in red in left panel) and data is visualized as a density scatter plot. The ~~colourbar~~colourbar represents the number of data points in each square.

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 (cv) 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 evaluation purposes. The cv results were strikingly similar as compared to the oob prediction (Table 2). The agreement was convincing which qualified the oob prediction as an appropriate way to quantify the generalization error of our RF model.

Table 2: Comparison of the RF generalization error quantified by the out-of-bag (oob) procedure and 10 fold cross-validation (cv) based on three metrics, i.e. coefficient of determination (R^2) root mean squared error (RMSE) and mean absolute error (MAE). The 1,900 additional observations were excluded for this evaluation.

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

Post training, the RF model was utilized to predict the depth to the shallow water table and the resulting map is shown in Figure 4. Regional patterns of the shallow water table were estimated as expected with deeper water tables in parts of the sandy meltwater plains in the western part of the domain and a water table that was overall close to the surface in the moraine landscape, as shown in Figure 1. Areas with low topography were generally exposed to a very shallow water table, which also corresponded to the conceptual understanding of the system. The 50 m spatial resolution provided a very detailed picture of the spatial patterns associated to the water table and the complex interplay of the covariates became apparent. This is shown on the basis of two zoom extents, highlighting urban areas, in Figure 4. The stream network and lakes are clearly ~~visual~~ visible with a depth of zero, which indicates that appending the additional observations to the training data resulted in the intended affect. The severity of the risk of groundwater induced flood events becomes apparent through the statistics of the RF map. The mean depth to the groundwater for a typical wintertime minimum event constituted 1.9 m for the entire modelling domain. Around 29 % of the domain was characterised by a depth to the shallow groundwater lower than 1 m and a depth of 50 cm or less ~~was evident~~ for 14 % of the area.

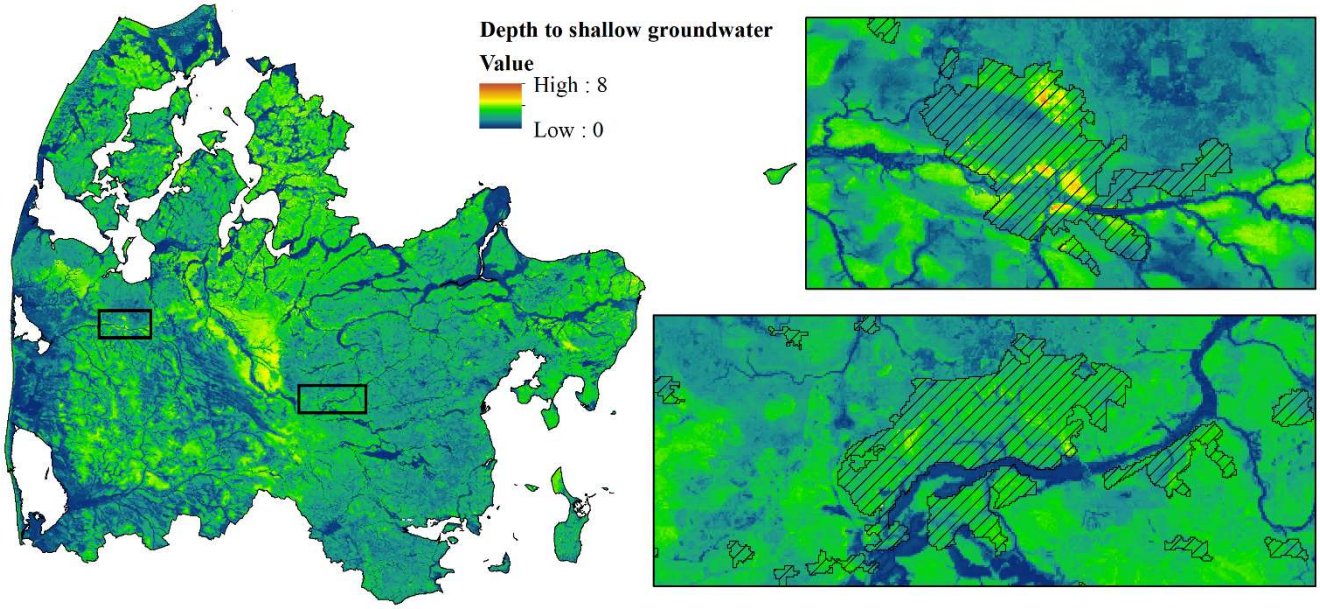

Figure 4: The resulting map of the depth to the shallow water table in 50 m grid resolution. The zoom extents highlight two urban areas. The top zoom displays the city of Holstebro and the bottom zoom depicts the city of Silkeborg. Urban areas are visualized in hatch signature.

3.2 Covariate Sensitivity ~~Analysis~~

Covariate sensitivity was analysed from two different perspectives, both using the concept of permutation accuracy. First, sensitivity was assessed for the trained RF model based on the decrease in R^2 of the oob prediction in consequence of permuting a covariate. For this, only the training dataset was incorporated which resulted in an overall covariate sensitivity score. Second, the sensitivity of the trained RF model was estimated individually for each simulation grid based on the absolute difference between a permuted and the original prediction. This approach gives the relative ranking of the most sensitive covariates for each simulation grid.~~Sensitivity analysis, i.e. assessment of covariate importance, was performed for the trained RF model and for the RF prediction.~~ Figure 5 shows the results for the former. The vertical distance to the nearest waterbody was the dominant covariate for the simulation of the shallow water table. A decrease of 60 % in performance was apparent when the variable was permuted. We found a direct relationship between the two variables, which highlighted that the shallow groundwater did not explicitly follow terrain variability. This resulted in a relatively deep water table at locations where the vertical distance is high and vice versa. The second most important variable in the trained RF model was the simulated water table by the national water resources model (DK-model), associated with a 15 % drop in performance when being permuted. The DK-model provides a typical minimum depth to the shallow water table at 500 m resolution for a 20 year reference period (1991 – 2010). This indicated that the DK-model could supply a valuable coarse trend to the RF model.

Figure 5 also quantifies the importance of physically related covariates. When permuted collectively, covariates associated to the topography resulted in a decrease of nearly 100 % in performance and thereby, the respective covariates formed the most important group. They were followed by covariates describing the waterbody relation (~70 % drop in performance) and geology related variables (60 %). As the vertical distance to the nearest waterbody relates to both, topography and waterbody proximity, it was included in both groups. The abovementioned results are based on the relative decrease in R^2 caused by permutation of covariates. In order to test if the resulting sensitivity ranking is metric dependent we conducted the same analysis based on the relative decrease in RMSE. We concluded that, in spite of varying absolute numbers, the same conclusion in terms of relative covariate sensitivity could be drawn and therefore, the results are not discussed further by this study.

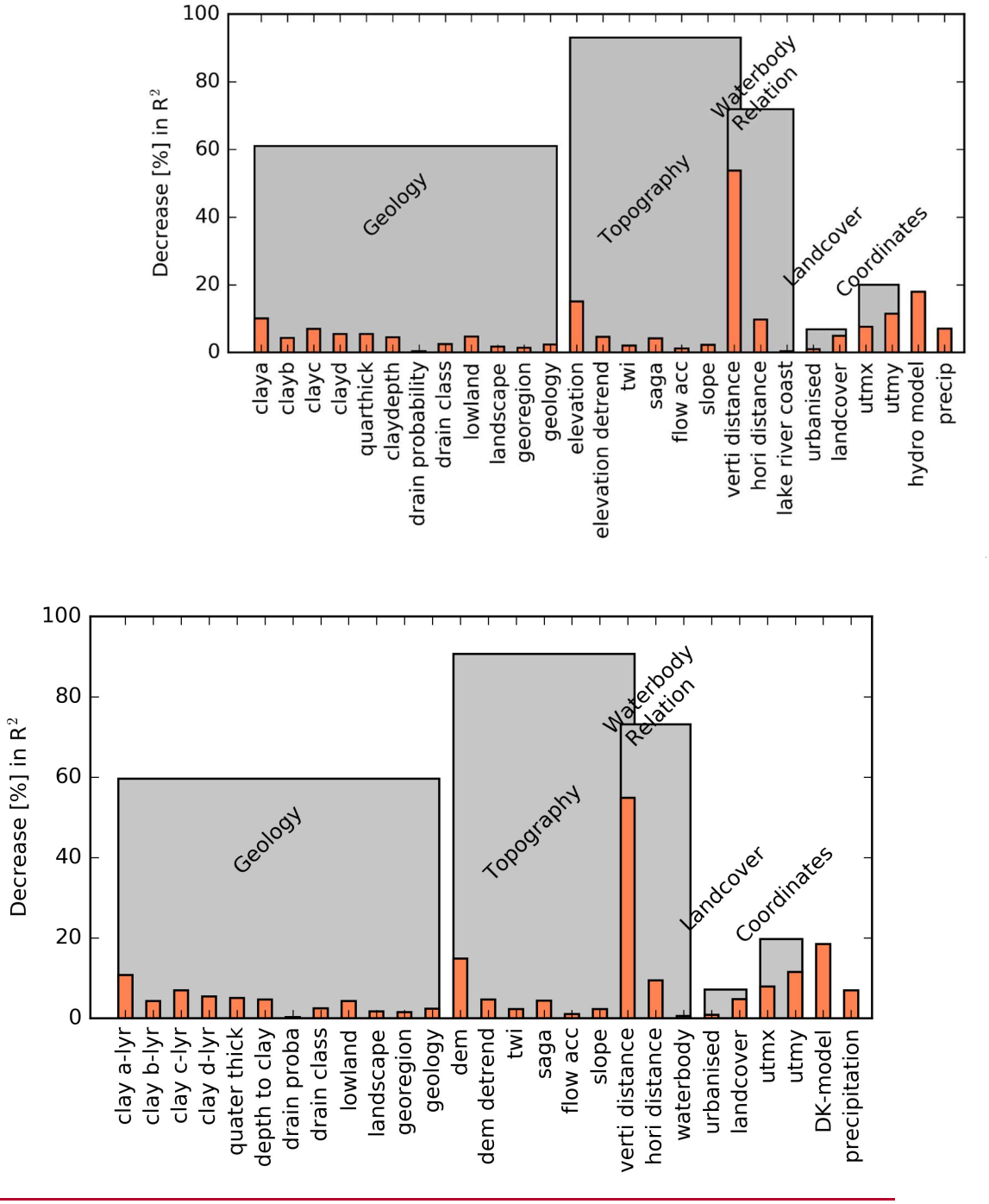

Figure 5. Variable importance of the trained RF model. The concept of permutation accuracy was implemented to quantify the decrease in out-of-bag performance R^2. Permutation was applied not only to single covariates (orange) but also to groups of covariates (grey). Covariates are further specified in Table 1.

The results from the spatial sensitivity analyses are presented in Figure 6. Figure 6 depicts maps of the top two most important covariates for the RF prediction. Covariates were permuted collectively following the groups presented in Table 1 and as applied in the sensitivity analysis of the trained RF model (Figure 5). Each covariate group was permuted 250 times to ensure that the difference to the original RF prediction converged at the individual grids. The simulated water table in the moraine landscape in the eastern part of the model domain was controlled by covariates related to the geology. Here topography is gently undulating and sediments are clay rich, which, in combination, resulted in a water table close to the surface with small-scale variability caused by geological heterogeneity. The second most important covariates in the moraine landscape were mainly the DK-model or the UTM coordinates. This underlined the complexity of the shallow water table in this landscape. The DK-model includes a comprehensive analysis of the entire system, taking the interplay between several factors, hydrogeology, topography, climate and others into consideration. In the RF model, coordinates provided the only possibility to assign uniqueness to a simulation grid, which was required in the moraine landscape to capture the complexity of the shallow water table. Topography was important at locations close to sea level or areas that were generally plane. Waterbody relations played an important role at location that were either very far away or very close to a waterbody. Data on the location of urban areas, which was contained in the land cover group, was rated important for urban areas with moraine soils. In such clayey conditions, the subsurface is often drained resulting in a deeper water table. Overall, the importance of the DK-model appeared to be very local and generally scattered across the domain, which underlined the relevance of this covariate, as it could provide coarse information at locations where the standard covariates fail at providing a meaningful generalisation.

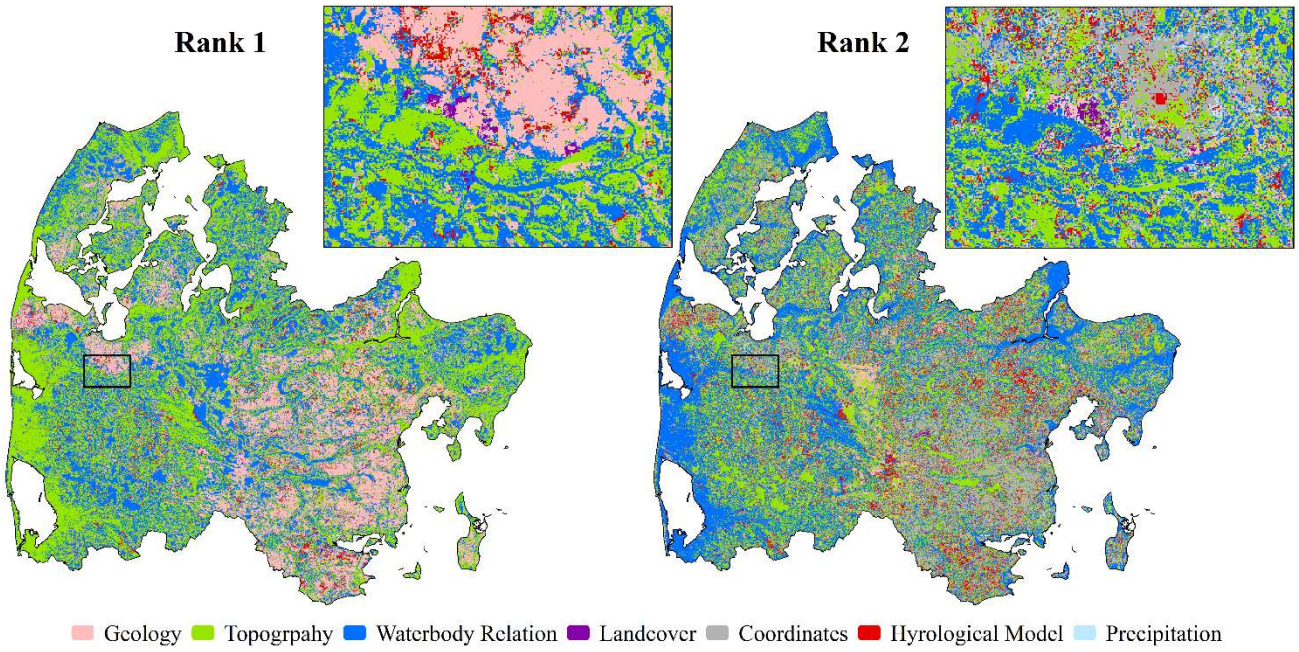

Figure 6: The results of the sensitivity analysis are shown for the most sensitive covariate group (Rank 1, left panel) and the second most important covariate group (Rank 2, right panel). The city of Holstebro is chosen as zoom for both maps.

3.3 Uncertainty Analysis

For the uncertainty analysis, we employed two methods, namely RFRK and QRF. For the first, the RF residuals were interpolated using kriging. Figure 7 shows the variogram model which was used in the kriging interpolation. The nugget was set to 0.26 m^2 and sill was defined as 1.02 m^2. An exponential variogram with a range of 700 m gave the most satisfying fit to the experimental semivariances calculated at 200 m lag distance.

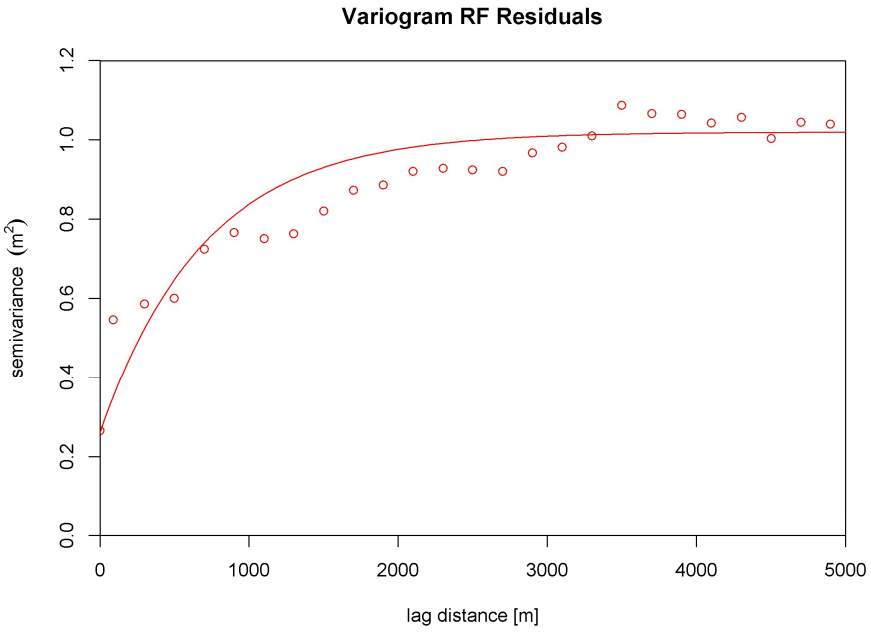

Figure 7: The computed semivariances for the RF residuals (circles), based on the oob prediction. The line expresses the fitted variogram model.

Figure 8 depicts the resulting uncertainty, which was expressed by the standard deviation for both of the applied methods. The RFRK employed all available data, ~15,000 wells and ~165,000 additional observations along streams, the coastline ~~coastlines~~ and in lakes in the RF residual interpolation. Opposed, the RF training dataset only contained 1,900 additional observations in order to have a balanced relationship of well data and additional observations in terms of data density. Following the predefined variogram, uncertainty was low in vicinity of an observation, which increased with distance until the sill value was reached. In this way, all grid cells with a distance to a well or additional observation larger than the correlation length of the variogram model exhibit no variation. Based on QRF, the derived uncertainty shows a different picture. Here the uncertainty was expressed as the standard deviation of the 1000 individual decision tree predictions at each ~~simulation~~ grid cell. In general the uncertainty estimated by RFRK was lower than QRF. In the western part, high uncertainties were in general associated to locations with a large depth to the shallow groundwater and vice versa for the QRF based assessment. However, the moraine landscape in the eastern part was characterised with an overall high uncertainty despite having an overall water table that is

close to the surface. Such physical dependencies that relate to the structure of the RF model were not captured by the RFRK approach, which purely reflected borehole proximity. ~~Therefore, we argue that the two approaches, to assess the uncertainty of the RF model, can be considered independent. Based on the concept of error propagation (eq. 3), we could take both sources of uncertainty, namely observation proximity and RF model structure, into consideration. Figure 8 presents the map depicting the combined uncertainty.~~ The mean uncertainty across the domain amounted to 0.92 m for RFRK and~~,~~ 1.68 m for QRF ~~and 1.93 for the combined map~~.

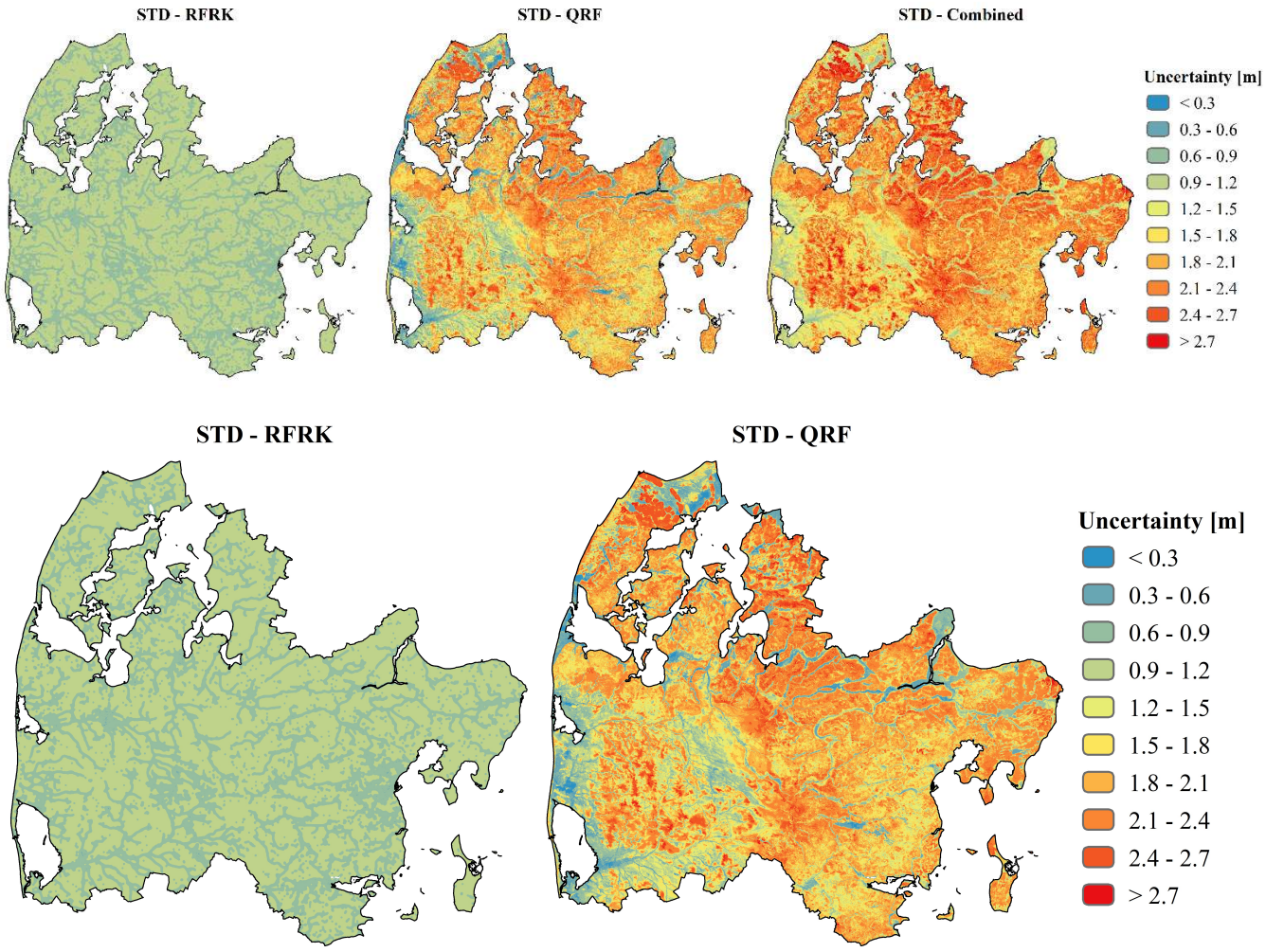

Figure 8: Two methods to quantify uncertainty of a RF model are implemented: RFRK (left panel) and QRF (~~middle~~ right panel~~). Under assumption of independence, they can be combined and results are depicted in the right panel.~~ For all maps, uncertainty is expressed as the standard deviation (STD).

4 Discussion

4.1 Training Dataset

In order to capitalize undersampled wells, this study utilized sinus curves, with amplitudes fitted to observations at wells with long timeseries according to their hydrogeological setting. Even though this step introduceds uncertainties, it was essential to generate a training dataset large enough to make robust predictions. Applying the same amplitude every year does did not distinguish between dry and wet years, which is was a clear limitation of the approach. The sinus curves described an average seasonal variation within a hydrogeological class of boreholes and are were thus not designed to reflect the variability of all boreholes within each class. Nevertheless, it iwass critical that the dataset used to train a RF model containsed a wide range of observations before the model is was able to generalize and make predictions. Along these lines, a training dataset can be expanded based on expert domain knowledge to capture otherwise underrepresented conditions (Koch et al., 2019). In this study, additional observations along streams, the coastline coastline and in lakes were appended to the training dataset with a depth to the water table of zero. In regions where the connection between surface and groundwater is generally good, like it is for this case in Denmark, the extent of surface waters can be considered a reliable proxy of the shallow water table. The additional observations used in this study guided the RF model to produce more reliable predictions. Initial tests without the additional observations resulted in surface waterbodies which were unconnected to the shallow groundwater system.

In unconfined sandy aquifers, we assume the elevation of the shallow water table to be more homogeneous than the depth of the shallow water table. This assumption may not apply for more complex geological settings, such as glacial tills, which cover a majority of the study area, where a secondary water table often follows the surface elevation. This motivated us to model water table depth instead of elevation, which was further supported by an initial test where a RF model was trained to predict water table elevation. The resulting water table elevation could easily be converted to depth by subtracting from surface elevation and results indicated a poorer performance compared to the RF model predicting water table depth.

The RF model was trained to a single event and thereby disregarding the temporal dynamics of the shallow groundwater system. Being designed as a simple screening tool, this can be considered an advantage; however, much of the complexity is not considered which is a clear a shortcoming of the proposed method. In future work, the sinus model can be replaced by the Danish national water resources model (DK-model) to correct the undersampled wells more physically based. Furthermore, this would allow a stringent extreme value analysis of the water table as well as climate change analysis. This was not included in this project as the resolution of the current DK-model is 500 m, which is too coarse to clearly differentiate temporal dynamics between hydrogeological settings at a scale reflecting the location of a well.

In the coming years, the Danish national water resources model (DK-model) DK-model will be updated based on recent hydrogeological interpretations and reconstructed in 100 m spatial resolution. This is expected to improve the predictability of the shallow water table and should then be utilized to update the RF model.

4.2 Random Forests Model

This study utilized the oob prediction to validate the performance of the RF model based on three metrics, namely coefficient of determination (R^2), mean absolute error (MAE) and root mean squared error (RMSE). The metric scores were overall very satisfying and in the range of what could be considered very acceptable in ~~traditional~~ groundwater flow modelling (Henriksen et al., 2003). These findings underpin the applicability of RF to model complex, non-linear variables with an accuracy ~~validity~~ that is difficult to obtain with ~~traditional~~ physically-based models. On the other hand, the ~~validation test~~accuracy assessment revealed ~~identified~~ a systematic bias of the trained RF model that was affecting wells with groundwater levels close to the terrain. The biased wells were predominately ~~placed~~ located in clayey moraine sediments, which indicated location specific shortcomings of the RF model. The geology of the moraine landscape is heterogeneous which impacts the hydrogeological setting and thereby also the shallow groundwater (Xiulan He et al., 2014, Xin He et al., 2015). At the current stage, the available national hydrogeological data do not possess the required spatial resolution to resolve the apparent heterogeneities adequately. Moreover, some of abovementioned wells are placed in confined conditions, which in combination with the heterogeneous geology may hinder good performance of the RF model.

Studying the covariate importance identified the water table simulated with the DK-model at 500 m resolution as the second most important RF input. These results were very promising as the applied RF framework forms a straightforward implementation of unifying machine learning and physically-based models. More precisely, RF buildt upon the coarse DK-model using high-resolution covariate information which ~~ensurined~~ ensured physical consistency.

Some covariates, e.g. drainage characteristics, topographic wetness index, were assigned an unanticipated low importance in the sensitivity analysis of the RF model (Figure 5). This may indicate covariate redundancy or the fact that the metric to quantify covariate importance, decrease in the coefficient of determination, is not very sensitive to the permutations which may result in changes of wells with a very shallow water table. As comparison, the RMSE instead of the R^2 was applied to quantify covariate importance, but this test did not provide any additional insights. Future work must systematically address the issues related to the choice of metric or the fact that certain parts of the distribution are more sensitive do different covariates ~~more systematically~~.

The resulting spatial resolution of 50 m provides a valuable screening tool for water management purposes. The risk of groundwater floods on agricultural fields or urban areas is typically very local and driven by small-scale variations of topography and geology. This makes high-resolution predictions inevitable in order to reliably tackle related challenges. At regional scale, the 50 m resolution would not be feasible with ~~traditional~~ numerical modelling, which emphasizes the versatile applicability of RF. Many covariates are available at finer resolution and, as computational power becomes more and more dispensable, RF predictions at even higher resolution are within reach. This development should also build upon current improvements of physically-based models, which are now already capable of providing results at resolutions in the range of hundred meters (Ko et al., 2019; Wood et al., 2011) and, thereby, such models could provide valuable trends, used as covariate in machine learning models.

This study proposed a novel approach to quantify covariate sensitivity of the simulation dataset, which results in a relative ranking of the most important covariates at grid level. This analysis provided physical insights on the driving mechanisms and, in general, the findings corresponded to the conceptual understanding of the hydrogeology in ~~that~~ the study area~~region~~. Such sensitivity maps are extremely valuable for both, the modeller and the stakeholders working with RF predictions. The former group can validate the physical consistency of the otherwise non-transparent black-box model and the latter will have a better understanding and ultimately also a greater acceptance of the predictions.~~. However, care must be taken, because the permutations must convergence in order to provide a robust sensitivity assessment.~~

4.3 Uncertainty assessment

This study assessed the capabilities of RFRK and QRF to estimate uncertainties associated to a RF model that predicts ~~ground~~water levels of the shallow groundwater system. Uncertainty was expressed by the standard deviation; alternatively, both methods could also be utilized to map upper and lower uncertainty bounds that represent certain confidence intervals. The key differences between the two proposed methods were as follows: (1) the uncertainty estimation of RFRK was in general lower than QRF and (2) the spatial patterns were diverging; RFRK reflected solely borehole proximity whereas QRF manifested a physical dependency of the uncertainty estimation. These findings are in line with recent comparison studies focusing on QRF and RFRK from the digital soil mapping literature (Szatmári and Pásztor, 2018; Vaysse and Lagacherie, 2017). Szatmári and Pásztor (2018) argue that RFRK based uncertainty estimations are limited because results do not depend on the data value and therefore, the method expresses an unconditional variance. This stringent assumption of homoscedasticity, i.e. constant error variance, could be unrealistic for variables where the variance behaves proportionally to the measured value (Hengl et al., 2018). Moreover, RFRK assumes that the RF prediction, which is used as trend, is certain and thus, the kriging variance only reflects the distance to the nearest observation. This assumption is too optimistic, as the uncertainty in the RF prediction is neglected. Once the training dataset is processed, RF disregards any uncertainties associated to the values of the target variable. In this study, uncertainties could originate from the applied sinus model used to transfer the observations to a typical wintertime minimum depth as well as the observations itself. In contrast, a physically-based hydrological model allows more transparency, as biased observations will be marked as outliers in the ~~validation~~ model evaluation~~step~~. However, a data driven model, as flexible as RF, will incorporate such outliers and thus biased predictions may arise.

As stated by Vaysse and Lagacherie (2017), QRF quantifies information of where a simulation point is located in the covariate space. In this way, QRF properly discriminates groundwater conditions of contrasted physical complexities of which some are better constrained by the training dataset than others. We argue that the RFRK shortcomings of assuming certainty in the trend prediction can be alleviated by the addition of QRF, which can capture the uncertainty of the RF model structure.~~. Following this hypothesis and assuming independence, error propagation can be applied (eq. 4). The combined uncertainty reflects both, uncertainty due to spatial proximity to the nearest observation, as provided by RFRK, and uncertainty induced by model structure, as quantified by QRF.~~ In summary, RFRK captures uncertainty related to the geographical space whereas QRF

describes uncertainties related to the covariate space. More work is needed to integrate these two sources of uncertainty into a single uncertainty quantification.

Reducing uncertainties can be achieved by collecting more observations and thus expanding the training dataset. Especially in the eastern part of the domain, which is characterized by a high clay content and a heterogeneous surficial geology, additional data would likely reduce the uncertainty. A measuring campaign in wintertime, when the shallow groundwater system is fully replenished, would be very beneficial to advancing the modelling capabilities. Additionally, a higher spatial resolution may contribute to an uncertainty reduction, as observations can be represented more uniquely by the covariates.

In more general terms, as the numbers of hydrological applications based on machine learning are vastly expanding, standards on how to conduct uncertainty analysis must be formalized in the same way this has been done for numerical modelling (Refsgaard et al., 2007). Ultimately, such a development conditions the stakeholder acceptance of machine learning results.

5 Conclusions

This study focused on using RF to predict a map that depicts the depth to the shallow groundwater at 50 m resolution for a typical wintertime minimum. More precisely, a minimum event that is expected to occur annually and poses risk of groundwater flooding affecting both, urban areas and agricultural fields. The regional map will be extremely valuable for water resources management. We draw the following main conclusions from our work:

1. RF is a versatile modelling tool with high accuracy that ~~allows to model~~enables spatial detail beyond the possibilities of ~~traditional~~, physically-based, numerical modelling. The depth to the shallow water table was modelled with a mean absolute error of 7~~9~~6 cm for an independent ~~validation~~ evaluation test.
2. Predictions from a coarse physically-based model that represent an overall trend of the water table can be utilized by RF as covariate. In this way, RF ensures physical consistency at coarse scale and exhausts high resolution information from topography, geology and other relevant variables. The DK-model at 500 m resolution was rated the second most important covariate in the trained RF model, indicating that this simple form of unifying machine learning and physically-based modelling has great potential.
3. The novel approach to assess covariate sensitivity for the prediction dataset goes beyond the standard applications where covariate importance is solely quantified for the training dataset. Results provide valuable insights on the spatial pattern of covariate sensitivity and can contribute to generate acceptability among end-users. The increased interpretability of the RF predictions can reassure modellers by comparing the derived sensitivity patterns with their conceptual understanding of the system.
4. In the general context of hydrological machine learning applications, more experience must be gained on how to properly quantify uncertainty. RFRK was found useful to assess observational proximity, but assuming certainty in the RF predications was regarded a shortcoming. This can be compensated by QRF, which is capable of addressing the uncertainty related to the structure of the RF model. ~~Simple, uncertainty propagation can be utilized to combine~~

~~both methods under the assumption of independence.~~ However, methods to take the uncertainties related to the observations itself and possible pre-processing of the training dataset are still lacking.

Acknowledgments

The work has been carried out with financial support granted by the Coast to Coast Climate Challenge project funded by the EU's LIFE programme.

References

Adhikari, K., Kheir, R. B., Greve, M. B., Bøcher, P. K., Malone, B. P., Minasny, B., McBratney, A. B. and Greve, M. H.: High-Resolution 3-D Mapping of Soil Texture in Denmark, Soil Sci. Soc. Am. J., doi:10.2136/sssaj2012.0275, 2013.

Adhikari, K., Hartemink, A. E., Minasny, B., Bou Kheir, R., Greve, M. B. and Greve, M. H.: Digital Mapping of Soil Organic Carbon Contents and Stocks in Denmark, edited by D. Hui, PLoS One, 9(8), e105519, doi:10.1371/journal.pone.0105519, 2014.

Ahmed, Z. U., Woodbury, P. B., Sanderman, J., Hawke, B., Jauss, V., Solomon, D. and Lehmann, J.: Assessing soil carbon vulnerability in the Western USA by geospatial modeling of pyrogenic and particulate carbon stocks, J. Geophys. Res. Biogeosciences, doi:10.1002/2016JG003488, 2017.

Asher, M. J., Croke, B. F. W., Jakeman, A. J. and Peeters, L. J. M.: A review of surrogate models and their application to groundwater modeling, Water Resour. Res., doi:10.1002/2015WR016967, 2015.

Banerjee, P., Prasad, R. K. and Singh, V. S.: Forecasting of groundwater level in hard rock region using artificial neural network, Environ. Geol., doi:10.1007/s00254-008-1619-z, 2009.

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., van den Hurk, B. J. J., Nearing, G. S., Pak, B., Peters-Lidard, C., Santanello, J. A., Stevens, L. and Vuichard, N.: The Plumbing of Land Surface Models: Benchmarking Model Performance, J. Hydrometeorol., 16(3), 1425–1442, doi:10.1175/JHM-D-14-0158.1, 2015.

Beven, K., Cloke, H., Pappenberger, F., Lamb, R. and Hunter, N.: Hyperresolution information and hyperresolution ignorance in modelling the hydrology of the land surface, Sci. China Earth Sci., 58(1), 25–35, 2015.

Biau, G. and Scornet, E.: A random forest guided tour, Test, doi:10.1007/s11749-016-0481-7, 2016.

Binzer, K. and Stockmarr, J.: soil, Kortserie- DGU, Danmarks Geol. Undersøgelse, 1994.

Böhner, J. and Selige, T.: Spatial prediction of soil attributes using terrain analysis and climate regionalisation, Göttinger Geogr. Abhandlungen, doi:10.1186/1471-2288-4-5, 2002.

Breiman, L.: Random forests, Mach. Learn., doi:10.1023/A:1010933404324, 2001.

Breuning-Madsen, H. and Jensen, N. H.: Pedological Regional Variations in Well-drained Soils, Denmark, Geogr. Tidsskr. J.

Geogr., 92(1), 61–69, doi:10.1080/00167223.1992.10649316, 1992.

Bricker, S. H., Banks, V. J., Galik, G., Tapete, D. and Jones, R.: Accounting for groundwater in future city visions, Land use policy, doi:10.1016/j.landusepol.2017.09.018, 2017.

Chaney, N. W., Van Huijgevoort, M. H. J., Shevliakova, E., Malyshev, S., Milly, P. C. D., Gauthier, P. P. G. and Sulman, B. N.: Harnessing big data to rethink land heterogeneity in Earth system models, Hydrol. Earth Syst. Sci., doi:10.5194/hess-22-3311-2018, 2018.

Close, M. E., Abraham, P., Humphries, B., Lilburne, L., Cuthill, T. and Wilson, S.: Predicting groundwater redox status on a regional scale using linear discriminant analysis, J. Contam. Hydrol., 191, 19–32, doi:10.1016/j.jconhyd.2016.04.006, 2016.

Daliakopoulos, I. N., Coulibaly, P. and Tsanis, I. K.: Groundwater level forecasting using artificial neural networks, J. Hydrol., doi:10.1016/j.jhydrol.2004.12.001, 2005.

Deutsch, C. V and Journel, a G.: GSLIB: Geostatistical software library and user's guide. [online] Available from: http://orton.catie.ac.cr/cgi-bin/wxis.exe/?IsisScript=orton.xis&method=post&formato=2&cantidad=1&expresion=mfn=072844, 1998.

Erickson, M. L., Elliott, S. M., Christenson, C. A. and Krall, A. L.: Predicting geogenic Arsenic in Drinking Water Wells in Glacial Aquifers, North-Central USA: Accounting for Depth-Dependent Features, Water Resour. Res., doi:10.1029/2018WR023106, 2018.

Fallah-Mehdipour, E., Bozorg Haddad, O. and Mariño, M. A.: Prediction and simulation of monthly groundwater levels by genetic programming, J. Hydro-Environment Res., doi:10.1016/j.jher.2013.03.005, 2013.

Fan, Y., Li, H. and Miguez-Macho, G.: Global patterns of groundwater table depth, Science (80-.)., doi:10.1126/science.1229881, 2013.

Ferguson, I. M. and Maxwell, R. M.: Role of groundwater in watershed response and land surface feedbacks under climate change, Water Resour. Res., doi:10.1029/2009WR008616, 2010.

Fienen, M. N., Masterson, J. P., Plant, N. G., Gutierrez, B. T. and Thieler, E. R.: Bridging groundwater models and decision support with a Bayesian network, Water Resour. Res., doi:10.1002/wrcr.20496, 2013.

Francke, T., López-Tarazón, J. A. and Schröder, B.: Estimation of suspended sediment concentration and yield using linear models, random forests and quantile regression forests, Hydrol. Process., doi:10.1002/hyp.7110, 2008.

Gleeson, T., Befus, K. M., Jasechko, S., Luijendijk, E. and Cardenas, M. B.: The global volume and distribution of modern groundwater, Nat. Geosci., doi:10.1038/ngeo2590, 2016.

Guo, P. T., Li, M. F., Luo, W., Tang, Q. F., Liu, Z. W. and Lin, Z. M.: Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach, Geoderma, 237–238, 49–59, doi:10.1016/j.geoderma.2014.08.009, 2015.

Hansen, M. and Pjetursson, B.: Free, online Danish shallow geological data, Geol. Surv. Denmark Greenl. Bull., 2011.

He, X., Koch, J., Sonnenborg, T. O., Jørgensen, F., Schamper, C. and Christian Refsgaard, J.: Transition probability-based

stochastic geological modeling using airborne geophysical data and borehole data, Water Resour. Res., 50(4), 3147–3169, doi:10.1002/2013WR014593, 2014.

He, X., Højberg, A. L., Jørgensen, F. and Refsgaard, J. C.: Assessing hydrological model predictive uncertainty using stochastically generated geological models, Hydrol. Process., 2015.

Hengl, T., Heuvelink, G. B. M. and Rossiter, D. G.: About regression-kriging: From equations to case studies, Comput. Geosci., 33(10), 1301–1315, doi:10.1016/j.cageo.2007.05.001, 2007.

Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., De Jesus, J. M., Tamene, L. and Tondoh, J. E.: Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions, PLoS One, 10(6), doi:10.1371/journal.pone.0125814, 2015.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M. and Gräler, B.: Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables, PeerJ, 6, e5518, doi:10.7717/peerj.5518, 2018.

Henriksen, H. J., Troldborg, L., Nyegaard, P., Sonnenborg, T. O., Refsgaard, J. C. and Madsen, B.: Methodology for construction, calibration and validation of a national hydrological model for Denmark, J. Hydrol., doi:10.1016/S0022-1694(03)00186-0, 2003.

Henriksen, H. J., Troldborg, L., Højberg, A. L. and Refsgaard, J. C.: Assessment of exploitable groundwater resources of Denmark by use of ensemble resource indicators and a numerical groundwater--surface water model, J. Hydrol., 348(1), 224–240, 2008.

Henriksen, H. J., Højberg, A. . L., Olsen, M., Seaby, L. P., van der Keur, P., Stisen, S., Troldborg, L., Sonnenborg, T. O. and Refsgaard, J. C.: Klimaeffekter på hydrologi og grundvand (Klimagrundvandskort), DANMARKS OG GRØNLANDS Geol. UNDERSØGELSE Rapp. (in Danish), 116, 2012.

Højberg, A. L., Troldborg, L., Stisen, S., Christensen, B. B. S. and Henriksen, H. J.: Stakeholder driven update and improvement of a national water resources model, Environ. Model. Softw., doi:10.1016/j.envsoft.2012.09.010, 2013.

Jankowfsky, S., Branger, F., Braud, I., Rodriguez, F., Debionne, S. and Viallet, P.: Assessing anthropogenic influence on the hydrology of small peri-urban catchments: Development of the object-oriented PUMMA model by integrating urban and rural hydrological models, J. Hydrol., doi:10.1016/j.jhydrol.2014.06.034, 2014.

Kahlown, M. A., Ashraf, M. and Zia-Ul-Haq: Effect of shallow groundwater table on crop water requirements and crop yields, Agric. Water Manag., doi:10.1016/j.agwat.2005.01.005, 2005.

Karlsson, I. B., Sonnenborg, T. O., Refsgaard, J. C., Trolle, D., Børgesen, C. D., Olesen, J. E., Jeppesen, E. and Jensen, K. H.: Combined effects of climate models, hydrological model structures and land use scenarios on hydrological impacts of climate change, J. Hydrol., doi:10.1016/j.jhydrol.2016.01.069, 2016.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N. and Kumar, V.: Theory-guided data science: A new paradigm for scientific discovery from data, IEEE Trans. Knowl. Data Eng., doi:10.1109/TKDE.2017.2720168, 2017.

Kidmose, J., Refsgaard, J. C., Troldborg, L., Seaby, L. P. and Escrivà, M. M.: Climate change impact on groundwater levels:

Ensemble modelling of extreme values, Hydrol. Earth Syst. Sci., doi:10.5194/hess-17-1619-2013, 2013.

Ko, A., Mascaro, G. and Vivoni, E. R.: Strategies to Improve and Evaluate Physics-Based Hyperresolution Hydrologic Simulations at Regional Basin Scales, Water Resour. Res., doi:10.1029/2018WR023521, 2019.

Koch, J., Stisen, S., Refsgaard, J. C., Ernstsen, V., Jakobsen, P. R. and Højberg, A. L.: Modeling Depth of the Redox Interface at High Resolution at National Scale Using Random Forest and Residual Gaussian Simulation, Water Resour. Res., 55(2), 1451–1469, doi:10.1029/2018WR023939, 2019.

Kollet, S. J. and Maxwell, R. M.: Capturing the influence of groundwater dynamics on land surface processes using an integrated, distributed watershed model, Water Resour. Res., 44(2), 2008.

Kreibich, H. and Thieken, A. H.: Assessment of damage caused by high groundwater inundation, Water Resour. Res., doi:10.1029/2007WR006621, 2008.

Larsen, M. A. D., Christensen, J. H., Drews, M., Butts, M. B. and Refsgaard, J. C.: Local control on precipitation in a fully coupled climate-hydrology model, Sci. Rep., doi:10.1038/srep22927, 2016.

Li, J., Heap, A. D., Potter, A. and Daniell, J. J.: Application of machine learning methods to spatial interpolation of environmental variables, Environ. Model. Softw., 26(12), 1647–1659, doi:10.1016/j.envsoft.2011.07.004, 2011.

Ließ, M., Glaser, B. and Huwe, B.: Uncertainty in the spatial prediction of soil texture. Comparison of regression tree and Random Forest models, Geoderma, doi:10.1016/j.geoderma.2011.10.010, 2012.

Lutz, S. R., Krieg, R., Müller, C., Zink, M., Knöller, K., Samaniego, L. and Merz, R.: Spatial Patterns of Water Age: Using Young Water Fractions to Improve the Characterization of Transit Times in Contrasting Catchments, Water Resour. Res., doi:10.1029/2017WR022216, 2018.

MacDonald, A., Hughes, A., Adams, B., Bloomfield, J., McKenzie, A. and Macdonald, D.: Improving the understanding of the risk from groundwater flooding in the UK, in Flood Risk Management: Research and Practice., 2010.

MacDonald, D., Dixon, A., Newell, A. and Hallaways, A.: Groundwater flooding within an urbanised flood plain, J. Flood Risk Manag., doi:10.1111/j.1753-318X.2011.01127.x, 2012.

Matheron, G.: Principles of geostatistics, Econ. Geol., doi:10.2113/gsecongeo.58.8.1246, 1963.

Maxwell, R. M. and Condon, L. E.: Connections between groundwater flow and transpiration partitioning, Science (80-.)., 353(6297), 377–380, doi:10.1126/science.aaf7891, 2016.

Meinshausen, N.: Quantile Regression Forests, J. Mach. Learn. Res., doi:10.1111/j.1541-0420.2010.01521.x, 2006.

Møller, A. B., Iversen, B. V., Beucher, A. and Greve, M. H.: Prediction of soil drainage classes in Denmark by means of decision tree classification, Geoderma, doi:10.1016/j.geoderma.2017.10.015, 2017.

Møller, A. B., Beucher, A., Iversen, B. V. and Greve, M. H.: Predicting artificially drained areas by means of a selective model ensemble, Geoderma, 320, 30–42, doi:10.1016/j.geoderma.2018.01.018, 2018.

Mutanga, O., Adam, E. and Cho, M. A.: High density biomass estimation for wetland vegetation using worldview-2 imagery and random forest regression algorithm, Int. J. Appl. Earth Obs. Geoinf., doi:10.1016/j.jag.2012.03.012, 2012.

Nearing, G. S., Mocko, D. M., Peters-Lidard, C. D., Kumar, S. V and Xia, Y.: Benchmarking NLDAS-2 Soil Moisture and

Evapotranspiration to Separate Uncertainty Contributions, J. Hydrometeorol., 17(3), 745–759, doi:10.1175/JHM-D-15-0063.1, 2016.

Nolan, B. T., Fienen, M. N. and Lorenz, D. L.: A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA, J. Hydrol., 531, 902–911, doi:10.1016/j.jhydrol.2015.10.025, 2015.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M. E. and Papritz, A.: Evaluation of digital soil mapping approaches with large sets of environmental covariates, SOIL, doi:10.5194/soil-4-1-2018, 2018.

Odeh, I. O. A., McBratney, A. B. and Chittleborough, D. J.: Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging, Geoderma, 67(3–4), 215–226, doi:10.1016/0016-7061(95)00007-B, 1995.

Pebesma, E. J.: Multivariable geostatistics in S: The gstat package, Comput. Geosci., doi:10.1016/j.cageo.2004.03.012, 2004.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É.: Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res., doi:10.1007/s13398-014-0173-7.2, 2012.

Refsgaard, J. C., van der Sluijs, J. P., Højberg, A. L. and Vanrolleghem, P. A.: Uncertainty in the environmental modelling process - A framework and guidance, Environ. Model. Softw., doi:10.1016/j.envsoft.2007.02.004, 2007.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N. and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566(7743), 195–204, doi:10.1038/s41586-019-0912-1, 2019.

Richey, A. S., Thomas, B. F., Lo, M. H., Famiglietti, J. S., Swenson, S. and Rodell, M.: Uncertainty in global groundwater storage estimates in a Total Groundwater Stress framework, Water Resour. Res., doi:10.1002/2015WR017351, 2015.

Rodell, M., Famiglietti, J. S., Wiese, D. N., Reager, J. T., Beaudoing, H. K., Landerer, F. W. and Lo, M. H.: Emerging trends in global freshwater availability, Nature, doi:10.1038/s41586-018-0123-1, 2018.

Rodriguez-Galiano, V., Mendes, M. P., Garcia-Soldado, M. J., Chica-Olmo, M. and Ribeiro, L.: Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain), Sci. Total Environ., 476–477, 189–206, doi:10.1016/j.scitotenv.2014.01.001, 2014.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. and Chica-Rivas, M.: Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines, Ore Geol. Rev., 71, 804–818, doi:10.1016/j.oregeorev.2015.01.001, 2015.

van Roosmalen, L., Christensen, B. S. B. and Sonnenborg, T. O.: Regional Differences in Climate Change Impacts on Groundwater and Stream Discharge in Denmark, Vadose Zo. J., doi:10.2136/vzj2006.0093, 2007.

Shen, C., Laloy, E., Albert, A., Chang, F.-J., Elshorbagy, A., Ganguly, S., Hsu, K., Kifer, D., Fang, Z., Fang, K., Li, D., Li, X. and Tsai, W.-P.: HESS Opinions: Deep learning as a promising avenue toward knowledge discovery in water sciences, Hydrol. Earth Syst. Sci. Discuss., 1–21, doi:10.5194/hess-2018-168, 2018.

Shiri, J., Kisi, O., Yoon, H., Lee, K. K. and Hossein Nazemi, A.: Predicting groundwater level fluctuations with meteorological

effect implications-A comparative study among soft computing techniques, Comput. Geosci., doi:10.1016/j.cageo.2013.01.007, 2013.

Stisen, S., Sonnenborg, T. O., Refsgaard, J. C., Koch, J., Bircher, S. and Jensen, K. H.: Moving beyond runoff calibration - Multi-constraint optimization of a surface-subsurface-atmosphere model, Hydrol. Process., (submitted), 2017.

Szatmári, G. and Pásztor, L.: Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms, Geoderma, doi:10.1016/J.GEODERMA.2018.09.008, 2018.

Tesoriero, A. J., Terziotti, S. and Abrams, D. B.: Predicting Redox Conditions in Groundwater at a Regional Scale, Environ. Sci. Technol., doi:10.1021/acs.est.5b01869, 2015.

Tesoriero, A. J., Gronberg, J. A., Juckem, P. F., Miller, M. P. and Austin, B. P.: Predicting redox-sensitive contaminant concentrations in groundwater using random forest classification, Water Resour. Res., doi:10.1002/2016WR020197, 2017.

Upton, K. A. and Jackson, C. R.: Simulation of the spatio-temporal extent of groundwater flooding using statistical methods of hydrograph classification and lumped parameter models, Hydrol. Process., doi:10.1002/hyp.7951, 2011.

Vaysse, K. and Lagacherie, P.: Using quantile regression forest to estimate uncertainty of digital soil mapping products, Geoderma, 291, 55–64, doi:10.1016/j.geoderma.2016.12.017, 2017.

Viscarra Rossel, R. A., Webster, R. and Kidd, D.: Mapping gamma radiation and its uncertainty from weathering products in a Tasmanian landscape with a proximal sensor and random forest kriging, Earth Surf. Process. Landforms, 39(6), 735–748, doi:10.1002/esp.3476, 2014.

Wang, F., Ducharne, A., Cheruy, F., Lo, M.-H. and Grandpeix, J.-Y.: Impact of a shallow groundwater table on the global water cycle in the IPSL land–atmosphere coupled model, Clim. Dyn., 50(9), 3505–3522, doi:10.1007/s00382-017-3820-9, 2018.

Winkel, L. H. E., Trang, P. T. K., Lan, V. M., Stengel, C., Amini, M., Ha, N. T., Viet, P. H. and Berg, M.: Arsenic pollution of groundwater in Vietnam exacerbated by deep aquifer exploitation for more than a century, Proc. Natl. Acad. Sci., 108(4), 1246–1251, doi:10.1073/PNAS.1011915108, 2011.

Wood, E. F., Roundy, J. K., Troy, T. J., Van Beek, L. P. H., Bierkens, M. F. P., Blyth, E., de Roo, A., Döll, P., Ek, M. and Famiglietti, J.: Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water, Water Resour. Res., 47(5), 2011.

Yoon, H., Jun, S. C., Hyun, Y., Bae, G. O. and Lee, K. K.: A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer, J. Hydrol., doi:10.1016/j.jhydrol.2010.11.002, 2011.

Youssef, A. M., Pourghasemi, H. R., Pourtaghi, Z. S. and Al-Katheeri, M. M.: Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia, Landslides, 13(5), 839–856, doi:10.1007/s10346-015-0614-1, 2016.

Zimmermann, B., Zimmermann, A., Turner, B. L., Francke, T. and Elsenbeer, H.: Connectivity of overland flow by drainage network expansion in a rain forest catchment, Water Resour. Res., doi:10.1002/2012WR012660, 2014.

Zipper, S. C., Soylu, M. E., Booth, E. G. and Loheide, S. P.: Untangling the effects of shallow groundwater and soil texture as drivers of subfield-scale yield variability, Water Resour. Res., doi:10.1002/2015WR017522, 2015.