Hydrology and Earth System Sciences

Open Access

Discussions

EGU

HESSD

Interactive
comment

Interactive comment on "Modelling of the shallow water table at high spatial resolution using Random Forests" *by* Julian Koch et al.

Katherine Ransom (Referee)

lockhart.katherine@gmail.com

Received and published: 24 June 2019

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.

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.

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?

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

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.

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.

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?

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.

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.

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.

In section 3.1 and Figure 3, are the very shallow water table points which were con-

sistently over-predicted the same additional points that were added (with 0 depth to water)?

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).

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

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?

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.

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?

Technical Corrections

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"

———————————————————

Printer-friendly version

Discussion paper