Authors' reply to review of "Machine learning-based downscaling of modelled climate change impacts on groundwater table depth" – 10.5194/hess-2022-122

Editor's comment	*Dear Dr Schneider,* *Your manuscript "Machine learning-based downscaling of modelled climate change impacts on groundwater table depth" has been subjected now to review by two reviewers. Unfortunately, a third review could not be obtained. The two reviews give in general a positive evaluation of your paper and recommend minor and major revision. I suggest major revision and additional review of the paper. The most important points to handle are: (i) address the issue of collinearity of explanatory variables/predictors used in the machine learning approach; (ii) clarify why extra points were used in the calibration and provide guidelines on a training dataset; (iii) improve the quality of some of the figures.* *In your answer to the main points and detailed comments, please indicate how comments have been handled exactly, indicating also whether text has been deleted and what the position of newly included text blocks is. I am looking forward to the new version of the paper.* *Best regards,* *Harrie-Jan Hendricks Franssen – editor*
Author's response	Dear Prof. Hendricks Franssen, Thanks a lot for handling our manuscript. We are grateful for the constructive feedback we received from the reviewers, which helped to improve the manuscript. We addressed their comments carefully as described in detail below. Best regards, Raphael Schneider on behalf of all co-authors

(line numbers in the authors' replies refer to the revised manuscript with marked changes)

Reply to Review by Anonymous Referee #1

Reviewer's comment	*The manuscript "Machine learning-based downscaling of modelled climate change impacts on groundwater table depth" by Schneider et al. presents a novel downscaling method which uses hydrological model simulation data at a coarse scale (500 meters) together with ancillary data (e.g. topography and hydrogeologic information) to derive indicators for groundwater changes for future climate scenarios at higher spatial resolution (100 meters). Model simulations at a scale of 100 meters for five selected catchments and five input data sets from different regional climate model simulations are used as training data for the downscaling algorithm which is based on the Random Forest method. Estimates of groundwater changes at high resolution are made by using hydrologic simulations at coarse scale (500 meters) with input from 18 regional climate model simulations. The downscaling method is verified with data from a high resolution (100 meters) simulation for one additional catchment.* *The topic of the paper is relevant to the hydrologic community as it describes an interesting possibility to provide stakeholders with high resolution information on potential changes in groundwater resources with an affordable computational cost. Generally, the paper is well written but there are a few issues that need to be clarified in my opinion.*
Authors' response	We thank the reviewer for the positive and constructive feedback to improve the manuscript. Below, we detail how we addressed the issues pointed out by the reviewer in the revision.

Reviewer's comment	*General comments:* *- The proposed downscaling method can be seen as a data-driven surrogate model for generating high resolution data out of the simulation results of the 500 meter model. This allows to avoid the computationally expensive direct simulation at this higher resolution, but adds some additional uncertainties and errors. In order to judge the quality and usefulness of these high resolution data, the user would still require some information on how the predictions improve when going from 500 to 100 meter resolution. Currently, the manuscript only provides information on how well the downscaling algorithm works but it does not describe the pratical benefits and improvements of the higher resolution. Hence I would suggest to add a paragraph (e.g. around line 123) that summarizes the main advantages of the high resolution model as inferred from previous comparisons of the low and high resolution model with observation data.*
Authors' response	The reviewer raises a valid point. When originally developing/calibrating the two versions (100m and 500m) of the model, the 100m resolution performed slightly better in terms of groundwater head performance (especially for shallow wells). However, we expect the 100m model to generally be better able to reproduce fine-scale variations of the uppermost groundwater level, as these are controlled largely by topography. And many of the relevant topographic variations will be smoothed out at 500m resolution, but remain visible in 100m resolution. These variations are hard to show with conventional groundwater observations, for example because some of the relevant areas such as river valleys are under-represented. However, we managed to show some of this benefit of the 100m by comparing satellite land surface temperature products (as a proxy for the shallow groundwater table) with modelled results across river valleys.
changes	Line 96ff: Added a sentence on the importance of fine-scale variations of topography and geology on the shallow groundwater table Line 126f: Mention improved shallow groundwater performance of the 100m HM over the 500m HM

Reviewer's comment	*In section 2.4.3 (lines 248-258) it is mentioned that additional points outside the five 'calibration' catchments were used in the calibration procedure of the algorithm to improve the robustness of the method. Can you explain in more detail what kind of robustness issues you detected? Do you have any explanation why these additional points were necessary although the five chosen 'calibration' catchments closely resempled the statistical properties for whole Denmark (Figure 2)? Which additional information did these 'dummy points' provide?*
Authors' response	With "robustness", we mean spatial transferability/performance on the spatial hold-out, i.e. using the algorithm outside the areas with training data. While it is true that the covariate space seems to be adequately covered by the training catchments, a random sampling of all of Denmark (the dummy/auxiliary points – we changed the term from "dummy" to "auxiliary"; see also next comment) still seems to be adding some covariate values/covariate combinations that inform the Random Forest regressor. Specifically, we for example could observe cases in spatial cross validation where the 100m downscaling result points (wrongly) in the opposite direction (positive or negative change in depth to groundwater table) without the use of auxiliary points. (In general, performance of a Random Forest algorithm or similar is not only determined by covering the covariate space, but also by covering the relevant combinations of the different covariates – a thought that was behind the development of the dissimilarity index by Meyer and Pebesma, 2021)
changes	Line 265ff: Extended/modified section 2.4.3

Reviewer's comment	*Additionally, the selection of additional calibration data through the 'dummy points' is not really in line with the argumentation in the rest of the paper which only refers to a calibration procedure with data from the five subcatchments. I would suggest to clearly state in all relevant parts how the calibration dataset was chosen (i.e. also mentioning the 'dummy points').*
Authors' response	The reviewer's point is correct. However, in case this was misunderstood, we want to point out that the dummy/auxiliary points originate from the coarse-scale resolution run of the hydrological model, so they did not require any additional runs of fine-scale hydrological models (i.e. they are taken from the to be downscaled variable). Maybe the wording "dummy points" was not ideal and caused some of the misunderstanding – we therefore changed the term "dummy points" to "auxiliary points" throughout the manuscript.
changes	Changed term "dummy points" to "auxiliary points" throughout the manuscript (and in Figure 3 and Table 2) Updated Figure 3 to include auxiliary points in the training data Updated caption of Table 2 (previously Table 1) to include the auxiliary points Line 262, 292: Included mention of the auxiliary points alongside the 100m training data.

Reviewer's comment	*Is it possible to provide guidelines on the size of the training data set? This would be an important information when applying the proposed downscaling method to other regions.*
Authors' response	That is a relevant question, but also a difficult one to answer. For a start, the necessary size of the training data depends a lot on the desired application. Are we only interested in (i) predictions within very limited areas/within the training catchments, or are we – as in the manuscript – interested in (ii) an algorithm that can be extrapolated beyond its training data? In case of (i), much smaller datasets than the one used here might be sufficient. In case of (ii), any possible answer probably is less related to a size of a training dataset, but rather to how well the training dataset covers the covariates (and covariate combinations) of the area to be extrapolated to (as also mentioned in the comment above when discussing "robustness").
changes	Line 426ff: Added some thoughts on size of training dataset/choice of training submodels for spatial transferability

Reviewer's comment	*Some plots are difficult to understand and need to be revised (see specific comments below).* *Specific comments:* *- Line 150: "...aggregated as described below." Please add the section number you are referring to.*
changes	Line 157f: Added "in section 2.3.1"

Reviewer's comment	*Line 154: It is not clear how the initial conditions were determined. Did you choose any random simulation time step between 1991 and 2100 as initial conditions or did you e.g. use the mean of this simulation period?*
Authors' response	We did use initial conditions from the actual simulation time step, i.e. from 01-01-2037 and 01-01-2067, respectively. The first four years following the simulation start were used as warm-up, but not included in the analysis.
changes	Line 163f: Added "using conditions from the same date as the respective simulation start"

Reviewer's comment	*Equation 1: Please make clear also through the notation that these statistics are calculated individually for each grid cell of the model.*

Authors' response	Agreed, we changed equation 1 and made it more clear in the surrounding text that this refers individually to each model cell. However, we (as explained in the added text) chose to omit the added "subscript g" in the remainder of the manuscript for ease of readability.
changes	Line 188ff: Adapted equation 1 and text as described above

Reviewer's comment	*Line 218: Please provide details on the "...differences between a historic dry and wet period,...".*
Authors' response	For this, we took the difference between a relatively dry historic period (the 12 consecutive years between 1990 and 2001; average yearly precipitation 817mm) and a relatively wet historic period (2004 to 2015; average yearly precipitation 852mm), and used the differences between respective groundwater levels in the same manner as the differences between future and reference periods.
changes	Line 231ff: Extended the explanation of this covariate as outlined above.

Reviewer's comment	*Line 222: Why is the 500 meter model output interpolated to 100 meters although this does not provide further information to the downscaling method? Is it a hard requirement of the algorithm to operate on equally sized vectors? Is there any explanation why the algorithm works better with interpolated TBDV data?*
Authors' response	Yes, the algorithm expects equally sized vectors, i.e. some kind of resampling from the coarse to the fine resolution has to be performed. Whether an interpolation (a simple bilinear interpolation in this case; not adding any data requirements or computational bottlenecks) is necessary or not remains unclear. However, in initial tests with non-interpolated data we experienced some artefacts from the edges of the 500m data in the 100m downscaled results.
changes	Line 240f: Added "[…]; using resampled TBDV without interpolation lead to visible artefacts at the 500 m grid boundaries in the 100 m outputs."

Reviewer's comment	*Line 392: Unit missing.*
changes	Line 426: Thanks for noticing; corrected to "100 m".

Reviewer's comment	*Figure 4: The scale break in the figure is a bit counterintuitive and misleading. I would suggest to show the different factors on a plot with the same scale (0 to 1) and add an additional plot (either separate or as an inset) with the second scale.*
Authors' response	Due to a comment of Reviewer #2, we completely redid Figure 4, which now includes not only the individual covariate importances, but also importances of perturbing entire groups of covariates. In this new version of Figure 4, the previously counterintuitive aspects are omitted.
changes	(modified Figure 4 due to comment by Reviewer #2)

Reviewer's comment	*Figure 5: Legends for the plots in the uppermost row seem to be missing. Generally, it is not readily clear with legend applies to which subplot.*
Authors' response	Correct, legends for the uppermost row are missing. This is on purpose, as the absolute values have no importance in this context; the two maps of relative topography and transmissivity in layer 1 are mostly shown to get an idea of how patterns in covariates influence patterns in the climate change impact.

changes	We added “map markers” a to f for each of the six maps in Figure 5, and now refer to those in the legend descriptions, the figure caption, and the manuscript text (line 343ff). (To ensure commonality, we applied similar changes to Figure 6, also adding map markers and the respective references to those; referred to in the text in line 363)

Reviewer's comment	*Figure 6: It is difficult to grasp what part of the verification data is shown in the different subplots (i.e. model input or output of the downscaling method). I would suggest to improve the figure headings and the caption text to guide the reader better through the figure.*
Authors' response	Good point; we made an effort to improve this by adding a column divider and column header to the two columns in the figure, making the distinction between the 5 RCM training ensemble and the 17 RCM full ensemble more clear. This is also reflected in the caption of Figure 6, as well as the manuscript text.
changes	Added column divider to Figure 6, as well as “map markers” (to ensure same commonality with Figure 6). This is reflected in the figure caption and manuscript text (line 363)

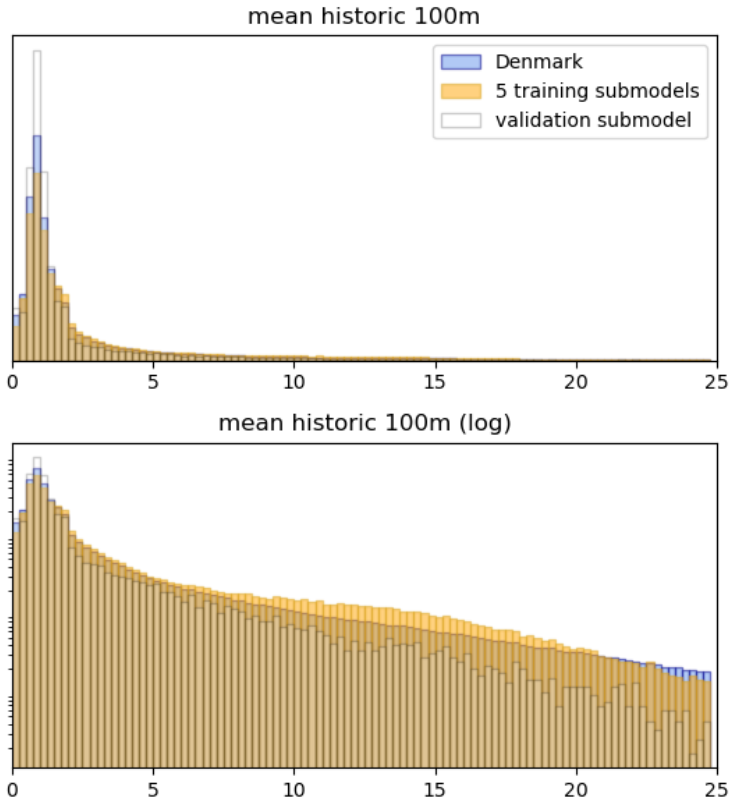
Reviewer's comment	*Figure 7: Please clarify the abbreviations in the figure, e.g. nf and ff. This might be guessed from the manuscript text but should also be made clear somewhere in the figure or the figure caption.*
Authors' response	The abbreviations are explained in section 2.3; however, this is long before Figure 7 in the manuscript, so we explain nf and ff again, as well as the difference between 500 m with interpolation (“500m HM intp”) and without. We consider the remaining abbreviations (mean, Q01, …) easier to remember and want to avoid repetition/an excessively long figure caption.
changes	Added “both in its original resolution and using bilinear interpolation to 100 m (“500m HM intp”)” and “for both near (nf, 2041-2070) and far future (ff, 2071-2100)” to the caption of Figure 7.

Reply to Review by Anonymous Referee #2

Reviewer's comment	*Schneider et al proposed a RF-based downscaling method to downscale changes in the simulated water table depth over Denmark from 500 m resolution to 100 m resolution under different future climate scenarios. The method was trained on data from five submodels that cover a wide range of geologic, topographic, and hydrologic variability occurring across Denmark, and validated on data from another submodel (VI). The results obtained by the proposed method outperformed 500m-resolution water table depth and its bilinear interpolation in showing the climate change-induced changes to the shallow groundwater table. The paper would be of interest to the hydrological community. Overall, it is well-written and the related questions are discussed thoroughly. However, I have the following concerns regarding the paper.*
Authors' response	We thank the reviewer for the positive and constructive feedback which contributed to improving the manuscript. Below, we detail how we addressed the concerns raised by the reviewer in the revision of the manuscript.

Reviewer's comment	*General comments:* *1. Traditional downscaling techniques downscale a product at a coarse resolution to the same product at a finer resolution. Here the authors used different statistics calculated from the coarse-resolution product (TBDV). Why didn't the authors directly use the 500m water table depth as a covariate here?*

Authors' response	Actually, we used the same statistics (mean, Q01, Q99, 1mex of changes to groundwater table) in 500m resolution (resampled/interpolated to 100m) as a covariate in downscaling to the respective output in 100m (i.e. again mean, Q01, Q99, or 1mex, respectively). Hence, if we understand the reviewer's comment correctly, our method is in that respect in line with "traditional downscaling techniques". On a side note: We also expect the proposed method to work with time-varying groundwater depth maps (as mentioned in the 2nd paragraph of the Conclusions).
changes	-

Reviewer's comment	*2. Which criteria did the authors use to select their validation submodel (VI. Aarhus Å/Aarhus)? From Fig.1, the submodel has a very shallow mean water table depth (0.5-2.5m). There are many areas in Demark having water table depth > 10 m, like the submodel V. I wonder if the selection of VI would give a biased conclusion for the RF validation.*
Authors' response	The five (plus one) submodels were picked from a set of ten submodels in total, which were part of the original model calibration (see section 2.1.3 and Henriksen et al., 2020a). These ten submodels already were chosen to be representative of hydrologic variations across Denmark. The submodels used in the downscaling training then were further selected from these ten based on their representativeness of relevant covariates (see Figure 2). Maybe this cannot be seen very well in Figure 1, but as can be seen in the plot below:  Submodel VI (the validation submodel) actually represents the Danish conditions quite well besides a moderate overrepresentation of shallow depths around 1m to 2m. The plot shows a histogram of the mean historic depth to the groundwater table, separately for Denmark, the five training submodels I to V and the validation submodel VI; with linear scale y-axis in the top plot (as in Figure 2), and log-scale in the bottom plot. Hence, while the reviewer correctly noted that there are significant areas with deeper groundwater tables in Denmark, we remain confident that those conditions are well covered by our training models (in particular submodel V) and also our validation submodel VI. On a side note: When it comes to vulnerability to climate change impacts, areas with currently shallow groundwater levels (within the first few metres of the surface) are of greatest interest.
changes	Changed Figure 2 to also include the histograms for the validation submodel Line 216ff: Adapted section 2.3.3 describing the submodel choice

Reviewer's comment

3. I really like the idea to study the importance of each covariate (feature) used in RF. I also think that determining the feature importance based on ML model performance is a feasible method. However, the authors may need to check the independence of their covariates before implementing such an approach. If two or more covariates are strongly correlated, perturbing one of them may not impact the ML performance, which leads to wrong results. I would like to know how the authors dealt with this issue.

Authors' response

Thanks. We agree with the reviewer, covariate correlation does affect feature importances. The covariates we used were already selected with covariate correlation in mind. Hence, covariate correlation is low for most of the covariates – see the table below for a matrix of pairwise covariate correlations (Pearson's R):

	topo	topod1-5	twi	drain_d	drain_tc	slope	reltop5	reltop2	trh_lay0	trh_lay1	trh_lay2	trh_lay3	trh_lay4	trh_lay5	trh_lay6	kh_mean_2-5m	kh_mean_5-10m	kh_mean_10-15m	kh_mean_15-20m	kh_mean_20-25m	tbdv	tbdv_hist	tbdv_hist_wd
topo	1.00	0.17	0.03	0.19	-0.08	0.05	0.26	0.19	-0.14	-0.02	0.08	0.02	0.06	0.03	-0.02	-0.07	-0.03	-0.02	-0.02	-0.01	0.11	-0.19	0.07
topod1-5	0.17	1.00	0.02	0.20	-0.05	0.03	0.69	0.69	-0.06	0.06	0.04	0.08	0.01	0.01	0.00	-0.04	-0.01	0.03	0.04	0.04	0.04	-0.19	0.02
twi	0.03	0.02	1.00	0.06	0.01	0.01	0.01	0.00	0.00	-0.01	-0.01	0.01	0.00	0.01	-0.02	-0.01	0.01	0.01	0.01	0.01	0.00	-0.01	0.00
drain_d	0.19	0.20	0.06	1.00	0.10	0.35	0.28	0.25	-0.13	0.04	0.07	0.10	0.06	-0.07	0.02	-0.02	0.00	0.01	0.00	-0.02	0.12	-0.19	-0.01
drain_tc	-0.08	-0.05	0.01	0.10	1.00	-0.11	-0.08	-0.06	0.33	0.13	0.08	0.08	-0.10	0.13	0.02	0.21	0.20	0.17	0.14	0.12	0.03	-0.06	-0.02
slope	0.05	0.03	0.01	0.35	-0.11	1.00	0.04	0.06	-0.09	0.03	0.11	0.07	0.08	-0.13	0.05	-0.01	-0.01	-0.03	-0.06	-0.08	0.10	-0.17	-0.01
reltop5	0.26	0.69	0.01	0.28	-0.08	0.04	1.00	0.89	-0.11	0.09	0.05	0.13	0.01	0.01	0.00	-0.06	-0.01	0.04	0.06	0.06	0.08	-0.28	0.02
reltop2	0.19	0.69	0.00	0.25	-0.06	0.06	0.89	1.00	-0.07	0.07	0.04	0.10	0.00	0.01	0.00	-0.05	-0.01	0.03	0.05	0.05	0.04	-0.22	0.02
trh_lay0	-0.14	-0.06	0.00	-0.13	0.33	-0.09	-0.11	-0.07	1.00	0.04	0.07	-0.05	-0.03	0.15	0.00	0.06	0.05	0.04	0.04	0.06	-0.06	0.02	-0.02
trh_lay1	-0.02	0.06	-0.01	0.04	0.13	0.03	0.09	0.07	0.04	1.00	0.16	0.31	-0.18	0.05	0.23	0.39	0.28	0.18	0.14	0.13	0.48	-0.47	-0.17
trh_lay2	0.08	0.04	-0.01	0.07	0.08	0.11	0.05	0.04	0.07	0.16	1.00	0.13	0.09	0.00	0.09	0.45	0.37	0.22	0.11	0.04	0.28	-0.31	-0.01
trh_lay3	0.02	0.08	0.01	0.10	0.08	0.07	0.13	0.10	-0.05	0.31	0.13	1.00	-0.23	0.01	0.24	0.43	0.53	0.58	0.54	0.43	0.46	-0.53	-0.15
trh_lay4	0.06	0.01	0.00	0.06	-0.10	0.08	0.01	0.00	-0.03	-0.18	0.09	-0.23	1.00	-0.15	-0.11	-0.22	-0.21	-0.18	-0.13	-0.07	-0.04	0.07	0.13
trh_lay5	0.03	0.01	0.01	-0.07	0.13	-0.13	0.01	0.01	0.15	0.05	0.00	0.01	-0.15	1.00	-0.04	0.17	0.16	0.16	0.14	0.13	0.01	-0.06	0.06
trh_lay6	-0.02	0.00	-0.02	0.02	0.02	0.05	0.00	0.00	0.00	0.23	0.09	0.24	-0.11	-0.04	1.00	0.12	0.12	0.12	0.08	0.03	0.12	-0.11	-0.15
kh_mean_2-5m	-0.07	-0.04	-0.01	-0.02	0.21	-0.01	-0.06	-0.05	0.06	0.39	0.45	0.43	-0.22	0.17	0.12	1.00	0.81	0.55	0.36	0.24	0.47	-0.52	-0.15
kh_mean_5-10m	-0.03	-0.01	0.01	0.00	0.20	-0.01	-0.01	-0.01	0.05	0.28	0.37	0.53	-0.21	0.16	0.12	0.81	1.00	0.71	0.44	0.26	0.47	-0.53	-0.17
kh_mean_10-15m	-0.02	0.03	0.01	0.01	0.17	-0.03	0.04	0.03	0.04	0.18	0.22	0.58	-0.18	0.16	0.12	0.55	0.71	1.00	0.70	0.44	0.41	-0.45	-0.14
kh_mean_15-20m	-0.02	0.04	0.01	0.00	0.14	-0.06	0.06	0.05	0.04	0.14	0.11	0.54	-0.13	0.14	0.08	0.36	0.44	0.70	1.00	0.70	0.32	-0.37	-0.10
kh_mean_20-25m	-0.01	0.04	0.01	-0.02	0.12	-0.08	0.06	0.05	0.06	0.13	0.04	0.43	-0.07	0.13	0.03	0.24	0.26	0.44	0.70	1.00	0.21	-0.26	-0.05
tbdv	0.11	0.04	0.00	0.12	0.03	0.10	0.08	0.04	-0.06	0.48	0.28	0.46	-0.04	0.01	0.12	0.47	0.47	0.41	0.32	0.21	1.00	-0.74	-0.26
tbdv_hist	-0.19	-0.19	-0.01	-0.19	-0.06	-0.17	-0.28	-0.22	0.02	-0.47	-0.31	-0.53	0.07	-0.06	-0.11	-0.52	-0.53	-0.45	-0.37	-0.26	-0.74	1.00	0.20
tbdv_hist_wd	0.07	0.02	0.00	-0.01	-0.02	-0.01	0.02	0.02	-0.02	-0.17	-0.01	-0.15	0.13	0.06	-0.15	-0.15	-0.17	-0.14	-0.10	-0.05	-0.26	0.20	1.00

We now mention the issue of covariate correlation more clearly in the manuscript (in section 2.4.4). Furthermore, we extended the feature importance analysis by a version where not only one covariate at a time is perturbed, but a whole group of (physically related) covariates is perturbed – similar to Figure 4 in Koch et al., 2019a. The comparison of the individual covariate importance with the covariate group importance shows no apparent issues with the covariate importance analysis due to covariate collinearity.

Accordingly, Figure 4 is changed, and the result section 3.1 is adapted.

changes

Lines 242f: Added information that the initial collinearity analysis lead to exclusion of some covariates
Lines 296ff: Added explanation of not only perturbing one covariate at a time for importance analysis, but also running it, where whole – physically related – groups are perturbed at a time.
Figure 4: Updated/changed, now also showing group-wise covariate importances
Line 327ff: Updated section 3.1 to include the discussion of the group-wise covariate importances

Reviewer's comment

4. Please improve the quality of the figures.

Authors' response	By that you mean the resolution and compression artefacts? In that case, we assume that the final article will be compiled in a different manner; the current quality issues are due to the pdf compiling.
changes	No changes for now to the figures' resolution/compression artefacts. Various changes to the contents, however, were made as a response to comments of both reviewers. (Figure 2, 4, 5, 6)

Reviewer's comment	*Specific comments:* *1. Line 102, Page 4: "referred to the provided literature". Which literature? (Abbott et al., 1986; DHI, 2020)? Please specify there.*
Authors' response	Here we mean the various references provided throughout sections 2.1 covering different aspects of the DK-model (subsurface parameterization, climate input).
changes	Line 104ff: changed to "For more details on model setup, input and parameterization, the readers are referred to the provided literature in the following two sections."

Reviewer's comment	*2. Line 112-114, Page 4: I am not an expert in hydrological model simulation, and I am a bit confused here. The authors mentioned that precipitation, temperature, and potential ET used for historic climate forcing to the DK-model HIP have various resolutions, 10 km or 20 km. However, in Line 105, they mentioned that all input data have a spatial resolution of 100 m. Therefore, did they downscale historical climate forcing data to 100 km or use them directly?*
Authors' response	Valid point. The climate forcing (at 10km/20km resolution) is interpolated to the model grid (500m or 100m).
changes	Line 118: Clarified this in the text

Reviewer's comment	*3. Climate models, Page 5: Can the authors clarify which 17 RCMs they chose and which 5 RCMs are used as a subset?*
Authors' response	We added the information, also explaining more on the choice.
changes	New Table 1 showing an overview of the 17 RCMs, including projected precipitation changes Lines 135ff: adapted manuscript text accordingly

Reviewer's comment	*4. Line 173, Page 6: Why did the authors use changes to the 1m exceedance probability? Can the authors explain the practical meaning of this statistic?*
Authors' response	Good question, relevant to be clarified. The threshold of 1m was chosen in connection with stakeholders and users of the data. Water levels closer than a certain threshold to the surface can create various challenges in agriculture, infrastructure and flooding. In this context, a threshold of 1m was considered relevant (also, the widespread tile drains in Danish agriculture are located at around 1m depth). The exceedance probability then indicates how frequently the respective threshold of 1m is exceeded, and how that probability changes with climate change.
changes	Lines 183ff: Added some explanation of the origin of the 1m exceedance probability

Reviewer's comment	*5. Line 235, Page 8: "RF is a supervised ML learning method; that means it requires training data". This statement is wrong in my opinion. Unsupervised ML methods also require training data. I think here the authors meant supplementary teacher signals that are used to guide the training process<s. In addition, ML is the abbreviation for machine learning. Please delete the extra "learning" here.*
Authors' response	The reviewer is correct; we were not precise with the choice of our words here. Thanks for also noting the typo.
changes	Lines 253ff: Reformulated to "RF is a supervised ML method, requiring labelled training data. Based on the training dataset, a RF regressor model learns about relationships between a set of covariates and the target (training) data values."

Reviewer's comment	*6. Please mark the locations of dummy points used in RF training in Fig.1 if possible.*
Authors' response	Due to the large number of dummy/auxiliary points (20,000), we think it is difficult to show them on the map. They are sampled randomly in space, from all of Denmark except for the areas covered by the training submodels (as described in section 2.4.3).
changes	-

Reviewer's comment	*7. Line 276, Page 9: I believe there should be Table 3.*
Authors' response	Correct, thanks for spotting the mistake.
changes	Line 303: Corrected (now it is Table 4, as we inserted a new Table 1)

Reviewer's comment	*8. Line 335, Page 11: which statistic does "the climate change-induced changes to the shallow groundwater table" indicate?*
Authors' response	Figure 7 gives an overview over all the eight TBDV (i.e. mean, Q01, Q99, and 1mex for both near and far future).
changes	Line 369f: Clarified this in the text (as well as the caption of Figure 7; see also the last comment of Reviewer #1)

Reviewer's comment	*9. Fig.7: Please explain the legends (e.g., 500m HM intp) in the caption.*
Authors' response	Valid point, now added
changes	Extended caption of Figure 7.