Reviewer 1

We would like to thank the reviewer for the careful and thorough reading of this manuscript and for the helpful suggestions, which certainly will improve the quality of this manuscript. Our response follows (the reviewer's comments are in *blue italics*). New responses in green.

General comment: *Review of "Incorporating interpretation uncertainties from deterministic 3D hydrostratigraphic models in groundwater models" by Enemark et al.*

When regional groundwater models are developed, an important step is to build a conceptual hydrostratigraphic model based on the geological information at hand. Conceptual hydrostratigraphic models are based on mapping the 3D juxtaposition of geological layers and translating these to aquifers and aquitards, which are subsequently populated with hydraulic parameters (conductivities, transmissivities, storage coefficients) and used to schematize the 3D-makeup of a groundwater flow and/or transport model. The mapping of geological layers is preferably done by expert geologists that combine their conceptual knowledge of the depositional or structural geological environment with in-situ borehole descriptions, outcrop information and geophysical data (e.g. gamma logs, EM measurements etc). However, since there is much room for interpretation, no two geologists will provide the same conceptual hydrostratigraphic model.

In this paper, the authors use a recently developed method to assess this "interpretation uncertainty" in hydrostratigraphic models to assess how the uncertainty about the layer boundaries between hydrostratigraphic propagates to the uncertainty in groundwater model outcomes. They compare this degree of uncertainty with the uncertainty that accrues from unknown hydraulic parameters (a more common analysis). Apart from demonstrating the method in an uncertainty analysis (focused on capture zone size and median travel time), the authors also show that the schematization uncertainty is important of little in-situ data are available and if the layers to be identified and mapped are thin.

This is a valuable paper that presents a nice approach that is worth being be picked up by the groundwater modelling community in order to extent their toolbox of approaches in uncertainty assessment.

I think this paper deserves being published in HESS subject to resolving the following issues.

Reply general comment: Thank you for the overall positive assessment of our study. The main points of concern raised by the referee will be addressed below in the corresponding specific detail comments.

MAJOR CORRECTIONS

Comment 1: *The Low-Frequency model is insufficiently explained in the paper. It may well be based on work by Madsen et al, but this paper needs be readable on its own. Particularly:*

1.1. How is the manual interpretation model constructed? Are the smooth lines between the interpretation points in Figure 2 actual kriged values? Was this a 2D-kriging per surface? What semivariogram was used? How does one make sure that the boundaries between layers do not cross or do cross in case of a presumed erosive surface? And how is this resolved? Is there a manul postprocessing?

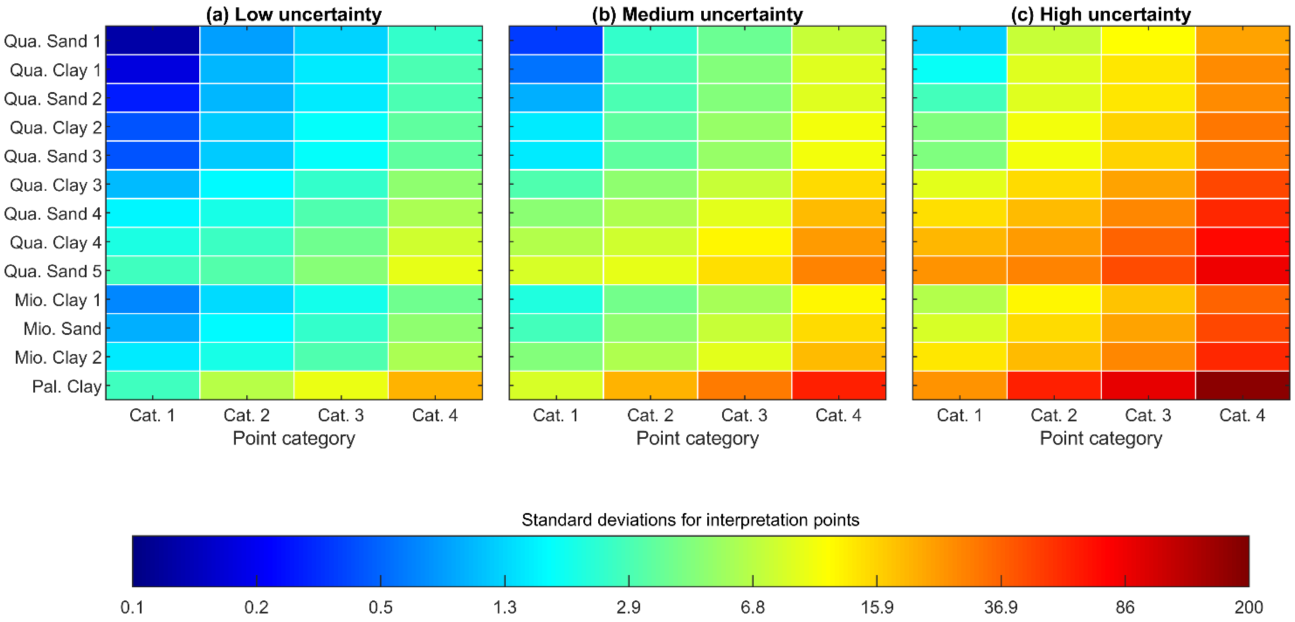
Reply 1: We will clarify this in the text. In short, the smooth lines shown on the figure is the layers that have been gridded from the interpretation points, using kriging. The semivariograms are adjusted for each layer surface during the geological modelling. The layers are adjusted using a layer adjustment tool in the modelling software (Geoscene3D), where the modeler defines the order of the layers.

Reply 1: To answer these questions, we have added extra information in the geological modelling section (section 2.1).

1.2. The simulation of vertical perturbations at the interpretation points is based on categories and then the standard deviation per category. A table with standard deviations should be given per case and per category.

Reply 1.2: We like the idea of a table showing the used standard deviations. However, we would like to split it into three separate tables as 3 scenarios x 14 layers x 4 categories results in a very large table with 168 entries. These three tables will represent the three different certainty cases. These tables will still contain 56 entries each. Thus, we would like to add these tables as supplementary material.

Reply 1.2: As requested, we have added the following figure as Supplementary Material S3, and referenced it in section 3.1.3, to visualize the tables described above.



1.3. Please be more specific about the nature of the LF model? How is it fitted? What are its equations? Are they smoothing splines? Or moving averages of a zero-error interpretation model interpolated with kriging between the randomly perturbed interpretation points? Or is it kriging with uncertain data? Perhaps a step-by-step procedure description would be helpful.

Reply 1.3: The LF-model is obtained by linear interpolation between interpretation points and then applying a smoothing kernel in a sliding window on the interpolated grid. The kernel width varies spatially to account for areas that contain steep transitions in the elevation. This is done by calculating the variance amongst the interpretation points within the window. If the variance is high, then the modeler has tried to convey large changes in the elevation and the smoothing becomes low to allow the LF-model to follow these structures. In the other case, a low variance of elevation amongst the points within the sliding window leads to more smoothing of the LF-model. The procedure is explained in-depth in Madsen et al. 2022, but all reviewers rightfully point out that this should be explained in more detail in the current manuscript. It was left out in the original draft to highlight the aspect of doing hydrological modelling on many hydrogeological model realizations rather than the generation of the hydrostratigraphy itself. Since this is clearly a weak point of the current manuscript as pointed out by all three reviewers, we will add some lines in the manuscript to summarize the steps taken to create the LF-model.

Reply 1.3: Section 3.1.2 has been updated to better describe the process and importance of constructing the LF-model for the hydrostratigraphy simulation.

1.4. How does the degree of smoothing takes account of spatially varying degrees of uncertainty? Is the level of smoothing spatially varying as well?

Reply 1.4: As explained above, the procedure for obtaining the LF-model takes spatial variability into account in terms of the curvature of the surface boundary. However, in the current setup level of smoothing is not affected by the uncertainty of the points. The uncertainty is instead utilized in the small-scale variability.

Reply 1.4: Section 3.1.2 has been updated to better describe the process and importance of constructing the LF-model for the hydrostratigraphy simulation.

1.5. It seems that with high uncertainty, the smoothing is more intense. But how does the LF model distinguish between actual locally higher spatial variability (while the data are certain and based on detailed borelogs) and spatial uncertainty?

Reply 1.5: In continuation of the previous reply, uncertainty is not related to smoothing in the LF-model. If so, this is purely accidental or that highly uncertain points maybe are interpreted more flat lying than in places where there is more certain information available. This would create more smoothing due to the low variance within the sliding window. This would probably make sense from an interpreter's point-of-view. In the revised version of the manuscript, we will address this apparent correlation between highly uncertain points and the variance in elevation between them.

Reply 1.5: We have added the following lines in section 3.1.2 to clarify this in the manuscript:

"In many places with low variance in elevation, high uncertainty is attributed to the interpretation points by the modeler. We attribute this apparent correlation to the fact that a modeler would likely not be willing to make bold structural interpretations in areas where there are no data to support such claims. Thus, the LF-model will indirectly carry less information in areas with high uncertainty, which is desirable, although the LF-model is based on the spatial structure of the interpretation points and not the uncertainty attributed to them."

Comment 2: *Section 3.2.3: here the 200 behavioural hydraulic parameter sets are selected with GLUE and assuming the manual interpretation model to be true. How is it guaranteed that these parameter sets are still behavioural for the other simulated realizations of layering? I understand that all combinations cannot be evaluated, but a few random hydrostratigraphies could be checked whether the 200 parameter sets are still close enough to be called behavioural?*

Reply 2: We agree that it cannot be guaranteed that the behavioral parameters in one model is behavioral in another model. The "behavioralness" of the realizations are illustrated in Figure 6. Realizations outside the red lines are non-behavioral. It is shown that almost 90 % of the realizations (between the 5th ad 95th percentile) are behavioral for the Low and Medium scenario while it is a bit less for the High uncertainty scenario. We will add a comment on this in the manuscript.

Reply 2: In section 4.2.1. we have added the following sentence to improve clarity: “It is shown that almost 90 % of the realizations (between the 5th and 95th percentile) are within the threshold values for the Low and Medium scenario while it is a bit less for the High uncertainty scenario”.

Comment 3: *The conclusion that with thick layers and lower uncertainty about the layer boundaries is small compared to the uncertainty from hydraulic parameters: To what extent is this conclusion dependent on model approach. How were the heads and fluxes simulated in these models: a quasi-3D aquifer-aquitard schematization or with a full 3D voxel model? In the first case, vertical fluxes in aquifers are ignored and this may underestimate the impacts of the thickness of a layer, particularly near wells. In this case it is also understandable that thickness (one order of magnitude variation) has much less impact than conductivity (with multiple orders of magnitude variation).*

Reply 3: The model we applied was a full 3D model in which the vertical discretization follows the hydrostratigraphic units. Vertical fluxes are considered. We realize that the statement in the conclusions may be a bit unclear and we will elaborate the argument in the revised version.

Reply 3: To clarify we have added the following sentence to the conclusion (section 6): “The applied groundwater model was a full 3D model in which the vertical discretization follows the hydrostratigraphic units.“

Comment 4: *I urge the authors to make the datasets (schematization, hard data, interpretation points and manual interpretation model) available.*

Reply 4: We agree that this is a good idea. Currently all geophysical data used for the manual modelling is available in the national Geophysical database of Denmark (GERDA) and all borehole information is available in the Danish borehole database (Jupiter). They can be accessed here: https://eng.geus.dk/products-services-facilities/data-and-maps/national-geophysical-database-gerda and https://eng.geus.dk/products-services-facilities/data-and-maps/national-well-database-jupiter. What is not currently available is the 3 sets of hydrostratigraphic models and the interpretation points used to generate them. It will however be possible to introduce both on the GEUS dataverse and be publicly available afterwards. We will do this for the forthcoming version of the manuscript and provide references.

Reply 4: The 3 sets of hydrostratigraphical models are now available online as well as the interpretation points. They can be accessed through the dataverse address:
https://dataverse.geus.dk/dataverse/EGEMODELS

MINOR CORRECTIONS

There are some small remarks and suggestions for improvements that I have put in the pdf attached.

Comment 5: *L28: something like a model cannot be uncertain; change to "subject to uncertainty".*
Reply 5: Thank you for the suggestion. We will fix it in the revision.

Reply 5: We have applied the suggested correction.

Comment 6: *L37: random is not the right word. unstructured would be better.*
Reply 6: Thank you for pointing out this mistake. We will fix it in the revision.

Reply 6: We have applied the suggested correction.

Comment 7: *L40: perhaps call this "interpretation"*
Reply 7: Thank you for the suggestion. We will fix it in the revision.

Reply 7: We have applied the suggested correction.

Comment 8: *L43: change to: propagate this to the uncertainty of the results of large-scale groundwater models.*
Reply 8: Thank you for the suggestion. We will fix it in the revision.

Reply 8: We have rephrased the sentence to "propagate this to the uncertainty of predictions from groundwater models".

Comment 9: *L70: Delete "particles".*
Reply 9: Thank you for pointing out this mistake. We will fix it in the revision.

Reply 9: We have applied the suggested correction.

Comment 10: *L76: which parameters? Please provide some examples. Hydraulic conductivity? Storage coefficient?*
Reply 10: Thank you for the suggestion. We will provide examples in the revision.

Reply 10: We have added hydraulic conductivity as an example.

Comment 11: *L93: First: Are the smooth lines between the interpretation points actual kriged values? And second: how does one make sure that boundaries do not cross or do cross in case of an errosive surface? And how is this resolved?*
Reply 11: The geologist provides a set of interpretation points to the geologic interpretation software GeoScene and the surfaces on the grid in between points are kriged or interpolated using some properties that can usually be set manually. Then for boundaries that crosscut there is functionalities that resolve these issues by the geologist setting a list of "erosive rules" to determine which layer's elevation should be adapted for both layers in the zone of overlap. The software then runs through all layers and post-processes each to make sure that there are no overlaps in the final 3D model. This practice was also used for each of the hydrostratigraphic realizations.

Reply 11: See answer to reviewer 1, reply 1.

Comment 12: *L112: delete "s" in represents.*
Reply 12: Thank you for pointing out this mistake. We will fix it in the revision.

Reply 12: We have applied the suggested correction.

Comment 13: *L166: Can you be more sp[specific about the nature of the LF model? How is it fitted? What are its equations? And how does the simulation algorithm then work? 1)Simulate elevations first; 2) then apply the smoothing? But should the smoothing not mean that residuals are also more or less random? No spatial correlation?;3) how does the smoothing distinguish between uncertainty and large spatial variability of layering.*
Reply 13: See reply 1.3.

Comment 14: *L224: How is it guaranteed that these parameter sets are still behavioural for the other simulated realizations of layering?*
Reply 14: See reply 2.

Comment 15: *L247: How is the entropy calculated?*
Reply 15: Say $f(m_i)$ represents the probability of outcome i out om N possible outcomes, then the entropy H, calculated with a log-base of *N,* is:

$$H = \sum_{i=1}^{N} f(m_i)\ log_N(f(m_i))$$

H will then be a number between 0 and 1, where 0 implies perfect knowledge (one outcome is certain), while H=1 will imply maximum entropy (maximum uncertainty), i.e. that all outcomes are equally probable, $f(m_i) = N^{-1}$.

Comment 16: *L332: How were the heads and fluxes simulated: quasi-3D or 3D. In the first case, vertical fluxes in aquifers are ignored and this may underestimate the impacts of the thickness of a layer. In this case it is understandable that thickness (linear scale) has much less impact than conductivity (order of magnitude scale).*
Reply 16: See reply 3.

Comment 17: *L393: This is a very nice way of representing this.*
Reply 17: We appreciate the positive feedback.

Reviewer 2

We would like to thank the reviewer for the careful and thorough reading of this manuscript and for the helpful suggestions, which certainly will improve the quality of this manuscript. Our response follows (the reviewer's comments are in *blue italics*).

General comment: *Dear authors, I read with interest your paper entitled "Incorporating interpretation uncertainties from deterministic 3D hydrostratigraphic models in groundwater models" which investigates the role of both hydrostratigraphic uncertainty and model parameter uncertainty on the prediction of groundwater models. The approach starts from an interpreted model which is perturbed by adding uncertainty (at different levels) on the boundaries between categories to produce 50 different realization. Then, each hydrostratigraphic realization is tested with a selected set of 200 model parameters combinations. It is concluded that the impact of hydrostratigraphic uncertainty is lower when the data density and reliability is large.*

I find the paper well written, scientifically rigorous and an important tool to characterize uncertainty. I recommend publication after taking into account the following suggestions.

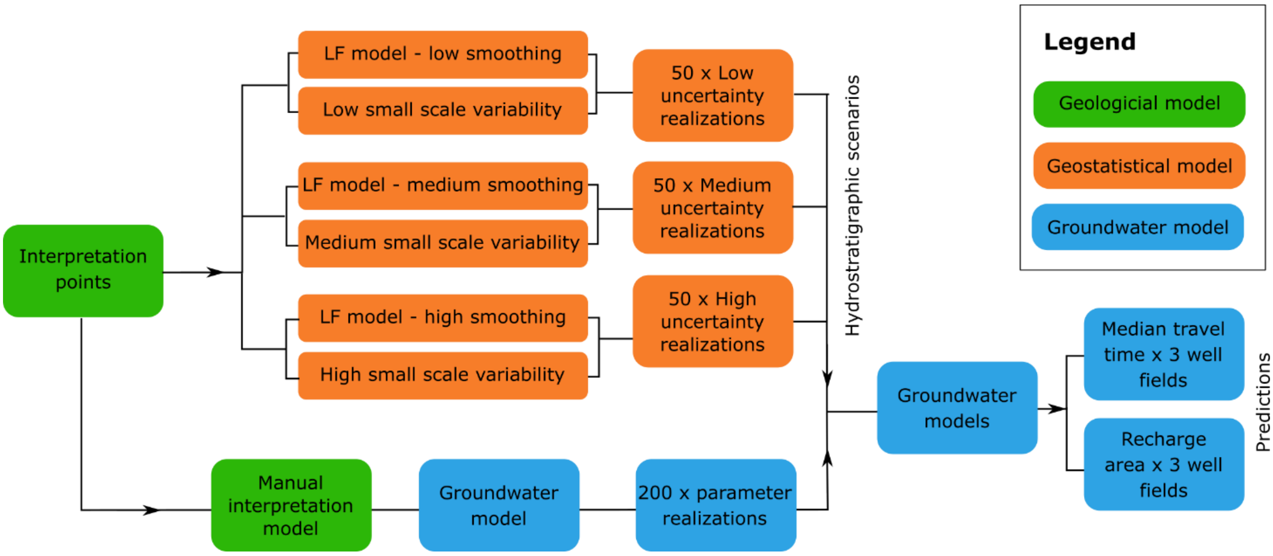
Reply General comment: Thank you for the overall positive assessment of our study. The main points of concern raised by the referee will be addressed below in the corresponding specific detail comments.

MAJOR CORRECTIONS

Comment 1: *It took me some time to understand that the paper would both investigate hydrostratigraphic interpretation and model parameters. In particular, the initial GLUE interpretation to select the model parameters was unexpected. I would therefore suggest to i) clarify this objective in the introduction, ii) to present a step-by-step workflow in the methodology, ideally accompanied by a figure, to clarify from the beginning the methodological approach.*

Reply 1: Thank you for the suggestion. To clarify the methodology, we will add a step-by-step workflow to section 3. In addition, we will make sure that the objective clearly stands out in the introduction.

Reply 1: To clarify the methodology, we have added a workflow figure to section 3 as well as a summary explaining the full workflow: "To evaluate the interpretation uncertainty of an initial interpreted hydrostratigraphic model, the hydrostratigraphic model is perturbed in three different uncertainty scenarios to produce 50 realizations in each scenario. In a groundwater model using the original interpreted hydrostratigraphic model, 200 parameter realizations are selected in a GLUE approach. The 200 behavioral parameter sets are applied using the hydrostratigraphic realizations. In the following, a detailed description of the methodology is provided."


Legend
Geologicial model
Geostatistical model
Groundwater model
LF model - low smoothing
Low small scale variability
50 x Low uncertainty realizations
LF model - medium smoothing
Medium small scale variability
50 x Medium uncertainty realizations
Interpretation points
LF model - high smoothing
High small scale variability
50 x High uncertainty realizations
Hydrostratigraphic scenarios
Groundwater models
Median travel time x 3 well fields
Recharge area x 3 well fields
Predictions
Manual interpretation model
Groundwater model
200 x parameter realizations

Comment 2: *I find the research context as currently presented in the introduction quite narrow. The field of uncertainty investigation in groundwater models is quite extended, and the introduction is rather written as an incremental step in the Danish methodology. Actually, what you propose could have much broader applications, as the same ideas could easily be applied to other geological modelling approaches (for example conditioned multiple-point geostatistics). Actually, I see some similarities with the work of Benoit et al. (2020, 2021) to simulate both hydrostratigraphic units and hydraulic conductivity uncertainty. In particular, your choice of selecting first 200 parameter distributions is justified by the desire to look at the marginal impact of hydrostratigraphy, but it ignores that zonation and model parameter likely interact (as likely illustrated by some of the rejected realizations you obtain), so that the posterior distribution for different scenarios are likely different (e.g., Hermans et al., 2015). Approaches that simultaneously simulate structural/scenario with model parameter uncertainty (possibly with intrafacies variability) could also be introduced/discussed.*

Reply 2: We agree that the introduction is written too narrow and much in the context the Danish research tradition within the subject area. In the revised version, we will broaden the introduction to reference better what has been tested in non-Danish cases and current state-of-the-art in uncertainty analysis of geological modelling in relation to hydrological models using the suggested references.

Reply 2: The introduction has now been rewritten and restructured. We believe it is now both more informative and more well-structued.

Comment 3: *I agree with reviewer 1 that the simulation approach to generate the hydrostratigraphic realizations should be better explained. Even if it is published in another paper, it is crucial for the current study and should therefore be included. In particular, the two step procedure (first category boundary, second LF model) could be illustrated with an example for each uncertainty level and corresponding simulations could be shown.*

Reply 3: As in the reply to Reviewer #1 we will expand the description of the LF methodology.

Reply 3: Section 3.1.2 has been updated to better describe the process and importance of constructing the LF-model for the hydrostratigraphy simulation.

MINOR CORRECTIONS

Below, I have a series of specific comments to further improve the manuscript.

Comment 4: *I wonder if using "deterministic" in the title is representative, as the uncertainty interpretation is actually based on stochastic simulations.*
Reply 4: We find the title is a fair representation of the content of the manuscript. The uncertainty realizations are indeed based on stochastic simulation. The starting point for the whole procedure is however a 3D static interpreted (deterministic) hydrostratigraphic model. The aim and challenge of this paper is to incorporate and propagate interpretation uncertainties in the resulting hydrological model.

Comment 5: *"typically assigned" suggests that this is standard practice, but only a reference in preparation is added. Do you have other references to cite ?*
Reply 5: Unfortunately, the only reference we have at this point is a guideline paper written in Danish and the mentioned paper, which is now a preprint. We will replace the word 'typically assigned' with 'may be assigned'.

Reply 5: The above stated correction has been applied.

Comment 6: *With "large-scale" do you mean the difficulty lies in upscaling the uncertainty? I am not getting the point, as if an uncertainty measure is given at the proper scale, it should not be more difficult that at the small scale.*
Reply 6: Thank you for pointing this out. In the revision "large-scale" will be removed from line 43.

Reply 6: We have applied the suggested correction.

Comment 7: *L67-69. You seem to make a difference between your approach starting from an interpreted model, and a stochastic approach that would start from a definition of prior probabilities. If conceptually different, isn't the end-result equivalent (a set of realizations), your interpreted model acting as a training image (e.g., Benoit et al., 2020).*
Reply 7: The resulting realizations can be seen as a probabilistic representation of the structural uncertainty (one can refer to it realizations from a structural prior). But the deterministic model is not used as a training image. We associate uncertainty to the layer interpretations and simulate the resulting hydrostratigraphic mode realizations stemming from that. This is quite different to using training images, or any other 3D type geostatistical model.

Reply 7: In section 3, we have added a short summary of the methodology as well as a figure showing the workflow. We hope that this has cleared up any misunderstanding.

Comment 8: *L91-98. I miss a better explanation here (only dealt with (partly) later in section 3.1.1) . From Figure 2, it seems that category 1 corresponds to wells, but it is unclear how the categories 2 to 4 are obtained (geophysics ?). How is the density of interpretation points chosen? For example, in a AEM image, one could have interpretation points all along the flight lines (sounding every few meters). See also main comment 3.*
Reply 8: We will expand the description in the revised version to increase the readability of the paper. The uncertainty categories are defined as follows:

1. Very certain interpretation based on certain borehole information.
2. Certain interpretation based on good quality unambiguous geophysical data and/or close to borehole data.
3. Intermediate uncertainty interpretation usually based on geophysical or borehole data of less good quality or ambiguous information.

4. Uncertain information usually based on interpretation of data of poor quality, extrapolated data, or no data at all.

The density of interpretation point is completely left the modeler. However, there are guidelines that one can choose to follow during interpretation. Usually for a good model there will be interpretation points for every 200-300 m along a profile. In areas with "pancake-like" stratigraphy one would probably use less interpretation points as the interpolation guides the surfaces in place quite comfortably. In other areas where the modeler wants to force a certain curvature of the surface and the geology is very complex, the need for interpretation points would increase.

Reply 8: See reviewer 1, reply 1.

Comment 9: *does or does not ?*
Reply 9: Thank you for pointing out this mistake. "does not" will be written in the revision.

Reply 9: We have applied the suggested correction.

Comment 10: *L159-160. It might be interesting to show an example (in Supplementary material?). Although this is not the topic of the article, it would help the reader to grasp what type of uncertainty we are talking about. For example, in absence of borehole, an AEM survey would not make the difference between sand layers of different ages, and thin layers might be missed. If uncertainty about a boundary is included, is the uncertainty about the presence of a boundary or not included in the different categories?*

Reply 10: This is indeed an interesting question. In such a case we totally rely on the conceptual understanding of the layer order and architecture as provided in the initial deterministic manual interpretation model. The initial model will never be better than the available data and the skills of the modeler. In these types of models there is not explicit information given on uncertainty of the presence of a boundary or not. This would be insightful to include and allow for a higher spatial variability than in the current setup. However, to obtain a meaningful relationship between the manual modeling, data resolution/availability and the certainty of which layers is present needs further investigation. Especially if one intends to superimpose this relationship on the simulations. We are working on this exact topic in related research.

Comment 11: *L161-163. This is again not the main topic of the paper, but from a geophysicist's perspective, the uncertainty about the presence of the layer is different from the uncertainty of the depth of the corresponding interface. I have no doubt that the clay is clearly visible on the geophysical inversion, however the uncertainty related to its depth depends on the regularization (smoothing), the depth of the interface (loss of resolution with depth) but also the discretization (geophysical inversion typically uses a grid whose cell size increases with depth). I would at least refer to the papers where this has been dealt with.*

Reply 11: Indeed, there is a difference in the uncertainty related to the presence of a specific layer and the uncertainty of its thickness and depth. We acknowledge that our primary crude approach to converting interpretation points relates to the discretization of the geophysical model and not the uncertainty related to how the geophysical model came about (regularization, measurement uncertainty, approximative physics, etc.). This would affect the results to include in the hydrostratigraphic model and was discussed in Madsen et al. 2022. We can try to incorporate some references we find relevant.

Reply 11: We extended the section of the paper to elaborate on the concept of "invisible" uncertainties of the geophysical model that will impact the interpretation uncertainty indirectly:

"E.g., the deep-lying Paleogene clay is well-resolved with electromagnetic methods due to its low electrical resistivity, thus having low interpretation uncertainty in the Manual Interpretation model (see e.g. Danielsen, 2003). Despite the comparatively low interpretation uncertainty, the chosen uncertainty should still reflect the uncertainties related to the processing and inversion of the geophysical data in the first place (Madsen et. al 2022, Viezzoli et al. 2013). Specifically for geophysical models, a commonly used quantifiable measure for the depth where the inversion result is no longer valid is provided by the Depth of Investigation (DOI) (Christiansen & Auken, 2012). Below the DOI, the attributed standard deviations should be among the highest possible in the study area. Above the DOI the quantification to standard deviations should factor in the resolution decrease of the inversion model with depth due to an increasing volume of the subsurface, which is being averaged over and the chosen regularization (Vignoli et al., 2015). Ideally, the standard deviation should also account for the possibility of inversion equivalences related to different parameterization of the inversion scheme (Høyer et al., 2014)."

Comment 12: *Section 3.1.2. Please see my main comment 3.*
Reply 12: Please see Reply 3.

Comment 13: *L197-198. I don't get the sentence. The travel time is surely dependent on the distribution of hydraulic conductivity in the area, while the zonation can (should) be an integral part of a calibration process.*
Reply 13: We agree that the clarity of this paragraph should be improved; we will rephrase it as follows in the revision: 'These predictions are chosen as they are not the calibration target but will be affected by the calibrated parameter zonation'.

Reply 13: We have applied the above stated correction.

Comment 14: *L222-233. See main comments 1 and 2.*
Reply 14: See replies above.

Comment 15: *Table 1. General Head and River conditions are mentioned in the table, but not in the description of the model and its boundary conditions. The outer boundary conditions (no flow?) and the recharge are not specified either. I guess the model is steady-state?*
Reply 15: We agree that the description of the boundary conditions should be improved. To address this issue, we will rephrase Section 3.2.2 in the revision as follows:

'The recharge to the water table is represented as a diffusive source with MODFLOW's recharge (RCH) package as a specified flux distributed over the top of the model. The well abstraction in the model is represented by the specified flux well (WEL) package. In all models regardless of the geometry of the model grid, the wells are set in the same layer to ensure model predictions can be compared i.e., the depth of the 210 abstraction wells may vary between realizations, but the layer and thereby the lithology will be the same in all models. The lakes and fjord in the southern part of the model area, is represented by the head dependent flux boundary General Head Boundary (GHB) package. To simulate both inflows to streams as well as subsurface tile drains and smaller ditches, the head-dependent flux boundary Drain (DRN) package is applied. The flux to the river cells is used as a calibration target'.

Reply 15: We have applied the above stated correction.

Comment 16: *There is no Enemark et al. (2021) in the reference list. Is it 2022?*
Reply 16: Thank you for pointing out this mistake. We will fix it in the revision.

Reply 16: We have applied the suggested correction.

Comment 17: *The transparency related to entropy is difficult to read on the figure because the initial color scale is made of nuances of the same colors (reddish, brownish). Maybe use a more diverse initial set of colors?*
Reply 17: We acknowledge that it may be difficult to recognize the fading due to the colors on the initial figure. The colors in the figure represent the typical colors for the sediments to make the figure more readable. However, we changed the colors of the Miocene deposits since the fading was too hard to see on the colors usually applied to these deposits. Thus, we have already experimented with the color selection and still this was the best version we could find.

Comment 18: *This is a steady-state model, isn't a convergence problem then just a matter of convergence criteria rather than a similarity issue with the Manual interpretation? For example because the solution would be further away from the initial state. Have you tried to use another solver or increase the number of iterations?*

Reply 18: We agree that the convergence issues are not a consequence of the dissimilarity with the Manual Interpretation model, but it may explain the difference in the convergence rates in the Manual Interpretation model and the realizations. We have not experimented with using different solvers, convergence criteria or more iterations. To improve the discussion about the convergence rates, we will revise L295-297 in section 4.2 as follows:

'In section 4.1 it was observed that the realizations of the Low uncertainty scenario are not necessarily more like that of the original Manual Interpretation model than the realizations of High uncertainty scenario, which may explain the difference in the convergence rates. The convergence rate is likely influenced by the model grids that are unique for each realization as it follows the layer elevations. The model grid is thereby influenced by the smoothing factor of the hydrostratigraphic model (Figure 3). The low smoothing factor of the Low uncertainty scenario allows larger changes in layer elevations than the high smoothing factor in the High uncertainty scenario. In areas where the layers are thin this may result in lack of lateral continuity between adjacent cells, which causes an inability to simulate flow between cells in the same layer'.

Reply 18: We have applied the above stated correction.

Comment 19: *Isn't it a problem of the set of parameter realizations that ate not adequate, because of the interaction between zonation and model parameters? See main comment 2.*
Reply 19: See reply 15.

Comment 20: *I guess the river flow is an average flow in the river. Since Modflow will simulate the average base flow to the river, isn't a large error expected (run-off component) in this case?*
Reply 20: In this area, the river flow is dominated by the baseflow component and the error introduced here is therefore considered to be limited.

Comment 21: *"Have" instead of "has"*
Reply 21: Thank you for pointing out this mistake. We will fix it in the revision.

Reply 21: We have applied the suggested correction.

Comment 22: *Delete "impact of"*
Reply 22: Thank you for the suggestion. We will fix it in the revision.

Reply 22: We have applied the suggested correction.

Comment 23: *L443-444. Other alternatives exist using for example simulation-based learning avoiding calibration (see recent review in HESS by Hermans et al. 2023 (section 3) and references therein or Thibaut et al. 2021 for a recent application to well head protection area – so a similar context). But as this is related to my own work, I am clearly biased and I let it to you to decide if this (and works from other) is relevant.*
Reply 23: Thank you for the suggestion. To convey that other alternatives exist than calibration of all geological realizations, we will add the following sentence to line 444 in the revision: "Another alternative is simulation-based learning to obtain the posterior distribution for the different scenarios (Hermans et al. 2023; Thibaut et al. 2021)".

Reply 23: We have applied the above stated correction.

Reviewer 3

We would like to thank the reviewer for the careful and thorough reading of this manuscript and for the helpful suggestions, which certainly will improve the quality of this manuscript. Our response follows (the reviewer's comments are in *blue italics*).

The paper "Incorporating interpretation uncertainties from deterministic 3D hydrostratigraphic models in groundwater models" addresses the importance of characterizing uncertainties of both the hydrostratigraphic model and the model parameters of the groundwater model. The topic is well presented and of high importance. Therefore, I recommend publication after the following points are addressed.

MAJOR CORRECTIONS

General Remarks:

Comment 1: *At first, I found it challenging to understand the procedure of the performed uncertainty quantification. For example that both the hydrostratigraphic and model parameter uncertainties are investigated. This should be clarified in several parts of the paper. For instance on p. 3 in l.71-76, three scenarios are mentioned. However, it should be clarified that not a single realization but multiple realizations are run per scenario.*
Reply 1: To clarify the methodology, we will expand Section 3 and explain the full workflow in more details.

Reply 1: See Reviewer 2, reply 1.

Comment 2: *There exists extensive literature about how to deal with uncertainties, especially in the field of geological modeling (e.g., Wellmann and Caumon, 2018), which needs to be added to the introduction to provide a broader perspective on this topic.*
Reply 2: We will broaden the introduction and relate our methodology to other geological modelling approaches. The suggested reference is highly relevant and a very enjoyable read. We purposedly avoided this topic as it was already covered in Madsen et al. 2022, but we acknowledge, as also pointed out by the other two reviewers, that we need to rewrite the current manuscript to be more stand-alone in terms of generating the hydrostratigraphic realizations.

Reply 2: The introduction has now been rewritten.

Comment 3: *I agree with the previous reviewers that the information provided about the low-frequency and manual interpretation model is insufficient and needs to be extended.*
Reply 3: See reply to Reviewer #1.

MINOR CORRECTIONS

Further Remarks:

Comment 4: *p.1 l.14: The authors talk about "the qualitative and subjective nature of uncertainty". In general, one distinguishes between epistemic and aleatoric uncertainties. While the statement is true for epistemic uncertainties it is not true for aleatoric uncertainties. So, the statement needs to be specified by explaining which type of uncertainties are addressed in the paper.*
Reply 4: In this study, we aim to characterize the aleatory part of the uncertainty associated with interpreting a hydrostratigraphic model. To avoid confusion, we will delete "subjective" from the sentence in the revision.

Reply 4: We have applied the above stated correction.

Comment 5: *p.2 l.42: A manuscript that is in preparation is cited. Please either publish that manuscript as a preprint and cite this preprint or use a different reference since the current reference is not available to the reader.*
Reply 5: The mentioned manuscript has since been published as a preprint and the reference will therefore be updated in the revised manuscript.

Reply 5: We have applied the above stated correction.

Comment 6: *p.2 l.42-43: Clarify which type of uncertainties the paper addresses (see also the first comment under further remarks)*
Reply 6: This paper addresses uncertainties related to the perceived uncertainty of the geologist' while producing the deterministic interpretation model. We will try to make this clearer in the text.

Reply 6: We have updated the sentenced to:

"Due to the qualitative and subjective nature of the uncertainty measure related to the perceived uncertainty of the geologist while producing the deterministic interpretation model, this information has previously been lost and hence, not incorporated in subsequent modelling, as groundwater modelling. "

Comment 7: *p. 5 Figure 1: It would be helpful to denote the profile lines with a,b, and c according to Figure 2. Such that these two figures can be better set in relation to each other.*
Reply 7: Thanks for suggestion, will do.

Reply 7: The figure has been updated with the suggestion.

Comment 8: *p. 5 l. 111: "The synthetic well field does exist in the real world … ". Should the formulation not be "The synthetic well field does not exist in the real world"?*
Reply 8: Thank you for pointing out the mistake. We will fix it in the revision.

Reply 8: We have applied the suggested correction.

Comment 9: *p. 10 l. 183-184: Why were 50 realizations chosen? Has a convergence test been performed?*
Reply 9: For computational reasons. The current setup scales badly because the number of model runs is a multiplication of both hydrological parameter realizations (200) and hydrostratigraphic models (50). According to the results obtained in Madsen et al. 2022, the minimum number of realizations to fairly represents the entropy of the hydrostratigraphy is 50. Thus, 50 realizations were chosen. However, more would have been better as we do acknowledge that in fairness it would probably be reasonable to have the same number of realizations of hydrostratigraphy as the number of hydrological parameter realizations. How to computationally optimize such a system is ongoing research in other related research projects.

Comment 10: *p. 10 l. 189: Where are the uncertainties of the medium scenario listed?*
Reply 10: We will make a comprehensive table with all the standard deviations used in all three scenarios in the revised version of the manuscript.

Reply 10: The table is constructed as a figure and is provided as supplementary material. See Reply 1.2.

Comment 11: *p. 11 Section 3.2.2: Provide the exact description and definitions of the boundary conditions and governing equations. It is not sufficient to list only the used packages.*

Reply 11: We agree that the description of the boundary conditions should be improved, see Reply 12 to Reviewer #2.

Comment 12: *p. 11 l. 215: Why was a random sampling strategy chosen and not a quasi-random strategy such as the Latin Hypercube Sampling (LHS) method? The LHS would have the advantage of better sampling the parameter space with few samples and avoiding the clustering of sample points as it often occurs for the random sampling method.*
Reply 12: In the manuscript, we have stated that random sampling has been applied, but we have actually applied a Latin Hypecube Sampling. The parameter values for the realizations are sampled from the prior parameter distributions using the Latin hypercube approach implemented using LHS class from the open-source Python framework pyDOE. The Latin hypercube designs obtained from pyDOE are transformed to uniform and log-uniform distributions using the values in Table 1 by applying the classes uniform and log-uniform, respectively, from the open-source Python. Thank you for pointing out this mistake. We will fix the mistake in the revision.

Reply 12: We have corrected the text in section 3.2.3 to indicate that we are using Latin Hypercube and random sampling.

Comment 13: *p. 11 l. 217: Why was a uniform prior chosen while all other considerations so far targeted normal distributions?*
Reply 13: For the parameter values in the groundwater model, all values within the range are considered equally likely and a uniform distribution have therefore been applied. For the interpretation points, the interpreted value is considered most likely, while values above or below the interpreted value with equal distance are considered equally likely and a normal distribution have therefore been applied.

Comment 14: *p. 11 l. 220: Provide details on how it was determined that the parameters are insensitive. Which type of analysis was used to get to this conclusion?*
Reply 14: A local sensitivity analysis was carried out in the Manual Interpretation model to evaluate the sensitivities of the parameters.

Comment 15: *p. 15 l. 294-296: The reasons why the solutions are not converging should be listed. Especially if the non-convergence is related to specific parameter ranges this might have a significant impact on the interpretation. Why is a trend of decreasing convergence problems observed with an increase in uncertainties? Would one not expect it to be the other way around?*
Reply 15: The non-convergence is not related to specific parameter ranges but rather to the individual model grids of the realizations. To improve the discussion on convergence, we will add the following sentences to section 4.2 in the revision:

'The convergence rate is likely influenced by the model grids that are unique for each realization as it follows the layer elevations. The model grid is thereby influenced by the smoothing factor of the hydrostratigraphic model (Figure 3). The low smoothing factor of the Low uncertainty scenario allows larger changes in layer elevations than the high smoothing factor in the High uncertainty scenario. In areas where the layers are thin this may result in lack of lateral continuity between adjacent cells, which causes an inability to simulate flow between cells in the same layer'.

Reply 15: We have applied the above stated correction.

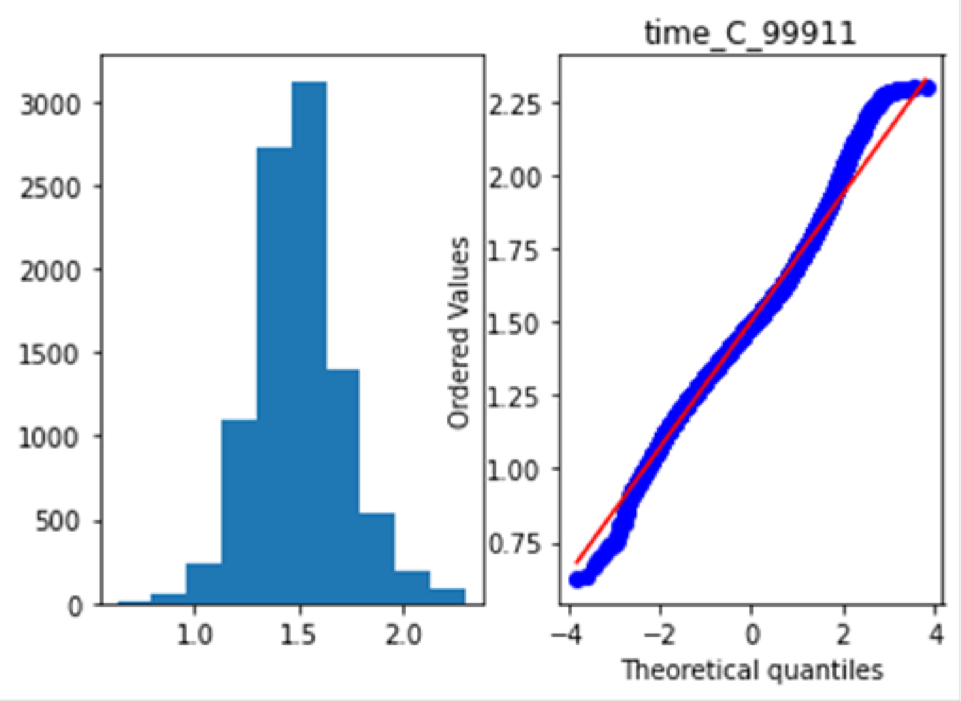
Comment 16: *p. 15 l. 298-299: It should be explained if and when why the analysis is still representative if in one case 46 % of the realizations are discarded and in the other two scenarios only 6 % or 1 %.*

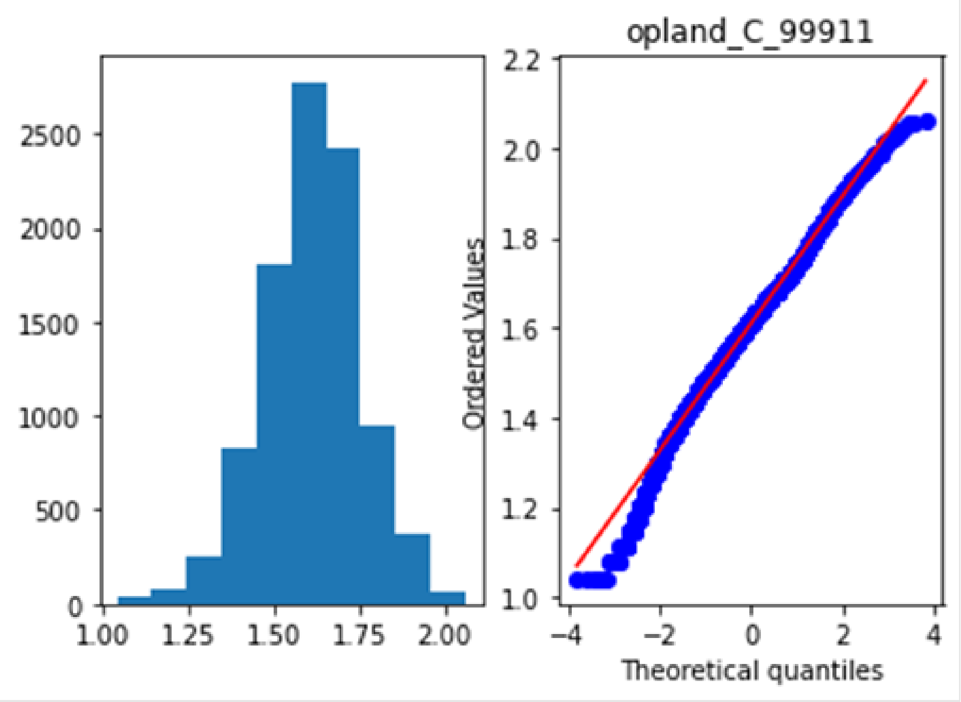
Reply 16: We agree that the results are not representative when 46 % of the realizations have been discarded. We have therefore indicated the results of this scenario in a parenthesis and with a dashed line in Figure 7. We will add a sentence to the caption of Figure 7 and line 403 to emphasize this in the revision.

Reply 16: We have added the following sentence to the caption of Figure 7: "For the Synthetic well field in the low uncertainty scenario half of the hydrostratigraphic realizations had to be discarded and the predictions are therefore placed in a parenthesis."

Comment 17: *p. 16 l. 320: The standard deviation is only a valid measure for normal distributions. Have, for instance, q-q plots been generated to show that the data follows a normal distribution?*

Reply 17: Thank you for pointing this out. Q-q plots have been generated and a few examples are shown below. As illustrated, deviation is seen around the edges of the distribution. However, considering that these areas have fewer data points, we acknowledge and tolerate this deviation.




opland_C_99911
2500
2000
1500
1000
500
0
1.00
1.25
1.50
1.75
2.00
2.2
2.0
1.8
1.6
1.4
1.2
1.0
Ordered Values
−4
−2
0
2
4
Theoretical quantiles