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.

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.

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.

MINOR CORRECTIONS

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

Comment 1: *I wonder if using "deterministic" in the title is representative, as the uncertainty interpretation is actually based on stochastic simulations.*
Reply 1: 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 2: *"typically assigned" suggests that this is standard practice, but only a reference in preparation is added. Do you have other references to cite ?*
Reply 2: 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'.

Comment 3: *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 3: Thank you for pointing this out. In the revision "large-scale" will be removed from line 43.

Comment 4: *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 4: 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.

Comment 5: *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 5: 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.

Comment 6: *does or does not ?*
Reply 6: Thank you for pointing out this mistake. “does not” will be written in the revision.

Comment 7: *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 7: 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 8: *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 8: 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.

Comment 9: *Section 3.1.2. Please see my main comment 3.*
Reply 9: Please see Reply 3.

Comment 10: *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 10: 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'.

Comment 11: *L222-233. See main comments 1 and 2.*
Reply 11: See replies above.

Comment 12: *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 12: 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'.

Comment 13: *There is no Enemark et al. (2021) in the reference list. Is it 2022?*
Reply 13: Thank you for pointing out this mistake. We will fix it in the revision.

Comment 14: *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 14: 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 15: *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 15: 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 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'.

Comment 16: *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 16: See reply 15.

Comment 17: *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 17: In this area, the river flow is dominated by the baseflow component and the error introduced here is therefore considered to be limited.

Comment 18: *"Have" instead of "has"*
Reply 18: Thank you for pointing out this mistake. We will fix it in the revision.

Comment 19: *Delete "impact of"*
Reply 19: Thank you for the suggestion. We will fix it in the revision.

Comment 20: *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 20: 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)".