Response to Reviewer #2 comments

We thank Reviewer #2 for taking the time to review this paper. We sincerely appreciate his/her insightful and constructive comments and suggestions. The response to the individual comments is given below. The original review is quoted in *italics*, whereas our response is given in **bold** font.

This paper presents a case study of multiple-point statistical simulation of sand/clay occurrence. The paper focuses on two aspects of multiple-point statistics: (1) 3D training image development and (2) different conditioning strategies to incorporate borehole data and geophysical data. This is a very relevant topic. Especially the construction of 3D training images is indeed still difficult. There are definitely some very interesting ideas in this paper, such as the different ways of using borehole data as hard or soft conditioning data. I would like to see, however, some more discussion on the following points:

In the title and aims of the paper, the authors stress the importance of realistic 3D training images. They write that they present a workflow to build a training image. The part about how they build the training image (section 5.4) is however very short. From this short description, it is not clear to me how exactly the training image was constructed. On what data is the TI based? On seismics or on the existing Miocene model? How were the shapes/geometry/position of the clay/sand featureds determined? What was the input in the Geoscene3D model? How exactly does this model work? What are the assumptions? Is this a manual or an automatic method? What are the "interpretation points" and where do they come from? They refer to a methodology in an earlier paper, but this paper uses completely different input data, i.e., airborne EM data?

On page 7, the authors write that "the results are evaluated and compared against the structures expected from the Miocene model"? How exactly? Was this model not already used as a basis for the TI? Is it then fair to use it again to evaluate the resulting structures?

In the manuscript, the TI characteristics are described in "5.4 Training Image", while the details of the strategy we have used during the TI development are provided at the beginning of the section "6. Results", and discussed thorough Fig.s 9 and 10. Along the paper (e.g., at: page 7, lines 15-22, page 9, lines 19-32, and page 10, lines 1-8), we emphasise the importance of unconditional realizations to assess the quality of the selected TI (actually, in combination with the used simulation algorithm). As it is shown by the comparison in Fig. 10, even small perturbations in the TIs (Fig. 9) affect significantly the associated realizations. After reading the Reviewer's comments, we feel that it would be probably more logical to move some of the material from page 7, lines 15-22 to the methodological section "5.4 Training Image". We will do this in the revised version of the manuscript. Moreover, we will add further details to the description of the TI construction: The first TI test was based on the existing 3D geological model (the Tønder model) covering an adjacent area. This first attempt was then manually adjusted

during several iterations based on the unconditional output (in the paper, for simplicity, only the initial and final TIs are shown). This iterative process stopped when the corresponding unconstrained realization was found satisfactory in terms of its ability to mimic the geological features we expect in the Miocene, across the study area. Those expectations about the geology are based on our prior geological understanding of the area, the available seismic lines, and the few existing deep boreholes. The entire procedure is manual (except, of course, the unconstrained simulation).

Regarding the "interpretation points", they relate to the interpretation of the geophysical data and the consequent construction of the associated (manual) 3D geological models (as described in Kristensen et al., 2015 and, shortly, in Fig. 5 in the manuscript). In the present workflow, in particular, they are used, within Geoscene3D, to efficiently modify the TIs. In fact, the interpretation points define the surfaces delineating the volumes to be populated with, for example, sand/clay voxels. By changing their locations and creating/deleting some of them, we could have a full control over the manual adjustments of the TIs.

Jørgensen et al., 2013 is mentioned simply as an example of research performed by using the voxel modelling tools available in Geoscene3D and specifically designed for (manual) 3D geological modelling.

The simulated structures are relatively simple and uniform. Is it really necessary to use MPS? If you would model sand/clay occurrence with a more simple method such as indicator kriging, you would probably get similar results? It would strengthen the paper if you can prove somehow that this relatively complex approach has significant advantages over simpler approaches.

We showed that even small details, which seem to be "irrelevant" in the TIs (that are capturing the available statistical information regarding the object to simulate), have actually significant impacts on the results (Fig.s 9 and 10).

Moreover, this is not meant to be a paper about the comparison of the performances of MPS as a tool to incorporate complex statistical information against other "simpler" approaches. This manuscript deals with the optimal strategies, within the MPS framework, to include the available data into the simulation as soft/hard conditioning and to build the most (geologically) effective TIs.

As a matter of fact, however, we believe that MPS' strength lies also in being quite intuitive: for many users, it is difficult to apply (and deeply understand), for example, indicator kriging and properly choose variogram models, while MPS allows the geologist to provide complex information in form of TIs, which is clearly a more intuitive (and, at same time, generally, a more effective) approach.

More in general, the conclusions of the paper are only based on visual inspection of the simulated clay/sand patterns. There is not objective or quantitative way of comparing the different results. For example in Figure 11: could you not use cross-validation or something similar to come to a more objective comparison of the different realizations? I also wonder how relevant the differences between the different realizations are, e.g. when you state that "the realizations showed a significant sensitivity to the TI". If you put these different realizations in a groundwater flow model, it is quite plausible that they all give similar results. It would be really interesting to using your geological model for some flow runs to see whether the different realizations based on different conditioning strategies really result in different groundwater flow patterns.

We agree that, in principle, it would be very interesting to investigate the results of flow models based on the different realizations. However, we feel that this would be out of the scope of the paper. Adding this kind of considerations would probably make the discussion lengthy and distract the reader from the main focuses of the paper, that are: 1) how to include, in the most effective way, the available data as simulation conditioning, and 2) how to build a TI that is capable to reproduce the geological features we want to see in the realizations. In particular, regarding the latter point, the ability to generate meaningful geological realizations should be seen in a general perspective and, so, relevant per se. Hence, even if, in this specific case, the differences between the two realizations in Fig 10 might be not "significant" for a flow model, the geological differences are evident, and, in principle, could impact the by-products generated by using the realizations. This paper is about the optimal practices to prepare the best possible stochastic geological inputs for subsequent applications.

Concerning quantitative analyses of the different realizations across the paper, maybe it would make clearer our point to provide, for the two realizations in Fig. 10, some measurements of, for example, the mean of the volume of the sand units. If this is found relevant, we will add it to the revised version of the paper.

In general, with respect to a quantitative analysis of the proportions variability of the different realizations discussed across the paper, we think that our point (i.e.: the kriged sand probability is effective in enforcing the proper spatial trend) is clearly demonstrated by the comparison between Fig 14b and Fig. 8. The two figures allow a voxel-by-voxel comparison between the soft conditioning distribution (derived from the boreholes) and the final corresponding realization. So, sincerely, we do not know which more quantitative argument we should use.

On the other hand, we agree that, in the text, we should emphasise more explicitly the relevance of this comparison.

The authors claim that in the study area many of the borehole records are of low quality. Why is that? In what sense are they of low quality? How can users in other study areas determine the quality of boreholes and decide whether they can treat the borehole data as hard or soft data?

Boreholes, like any other kinds of data, are affected by noise. The level of noise determines the quality/reliability of the measurements (i.e., the quality/reliability of the boreholes). Many factors impact the quality of boreholes. Just to mention a few of them: (i) the drilling methods (e.g.: hydraulic-rotary drilling and air core drillings. In some cases, for example, the finer sediments can be flushed out and the driller could potentially misinterpret, as more sandy, a clay layer); (ii) the drilling purpose (sometimes, if the goal is to reach a specific target, the lithological description can be quite poor since it is not a priority); (iii) the age of the boreholes (nowadays, in Denmark, samples are collected systematically every meter. For sure, this was not the case few years ago); (iv) the presence of simultaneous wireline logging data (these kinds of ancillary information make the geological interpretation definitely more certain). During the geological modelling phase, a skilled geologist goes through all the borehole records and verifies all these different pieces of information and check for inconsistencies.

In the paper, we mentioned that, for example, we use the seismic data as hard data because they were considered highly reliable and the scale of the structures they were able to delineate was comparable to the scale of the simulated outputs. Clearly, this is not true for the boreholes.

The results of the paper are based on visual comparisons of individual realizations using different conditioning strategies. It would be really interesting to see multiple realizations for each conditioning strategy to see the variability and uncertainty of different realizations.

We agree with the reviewer that investigating the variability and uncertainty would be interesting. However, we believe that this would be out of the scope of the present paper.

On page 9, the authors write that realization rarely have the same spatial variability as the training image. I find this a really strange statement. The idea of MPS is that you produce realizations with similar spatial patterns as the TI. If the realizations do not show similar patterns, this usually means that the parameters have not been chosen optimally or that the TI was for example not large enough. They also claim that therefore a TI should be chosen together with a specific MPS algorithm? If find this really strange. In my experience, all MPS algorithms can produce realizations with similar patterns to the TI given that they are used in the right way. If the authors really want to claim that different MPS algorithms have different capabilities in reproducing patterns, they should show a comparison of the different algorithms.

The meaning of what we wrote at page 9 concerns the fact that, while, ideally, the effects of the TI should not depend on the algorithm choice, these is not true in practice. The different implementations, and the different parameters that can be tuned during the simulation, make the realizations algorithm-dependent. It would not make much sense to compare the realizations generated with different algorithms, not only because different algorithms have different

parameters to be set, but also because the point of the paper is that the TI must be developed, no matter what, by studying the unconstrained realization and its accordance with our geological expectations.

Basically, we claim that the TI and the simulation algorithm are all interdependent elements structuring the final models, and, as such, the evaluation should take place a-posteriori on unconditional realizations.

On page 12, the authors write that "probabilistic models need to be developed and refined in order to utilize the multiple realizations and the uncertainty the represent". There are however many methods available and applications of MPS using multiple realizations to assess the uncertainty. The methods to do this are available but the authors have chosen to work with single realizations.

Writing that, we meant that applications and methods that are making use of the full potential (for example for geological modelling) of probabilistic models (with their large number of realizations and information regarding the uncertainty) are still under development. Those lines in the original manuscript were not about the possibility to assess the uncertainty via analyzing the realization ensemble.

The reasons why we decided not to investigate the uncertainty are largely discussed before in this document. And, accordingly to the future editor's suggestions, we can make these reasons more explicit in the next version of the paper.

The paper is clear and well written. The figures are of good quality.

P4, line 24-25: "along extended profile": error in grammar of sentence?

We thank once more the reviewer for the positive comments, especially because we invested a lot of time in the preparation of good quality and, hopefully, clear figures.

We believe that "along extended profile" reflects what we mean.