Response to referees, hess-2017-413

Firstly, the authors would like to thank the two anonymous referees for taking their time to read the manuscript and providing detailed and constructive comments. The comments, questions and suggestions are addressed in the following sections.

***AC:** Author Comments
***RC:** Referee Comments

Response to anonymous referee #1

General comments:
'*The paper applies and compares three multiple point statistics (MPS) methods (snesim, DS and iqsim) for hydrostratigraphic modelling using geological and geophysical data. This research is very relevant as (1) three MPS methods, including very recent methods, are compared to evaluate the advantages and disadvantages of each method which is very useful for users that want to select one of the different available MPS methods and (2) since these methods are all applied on a real-world case with realistic geological complexity and data availability.*

RC1: '*The authors first apply the three MPS methods on their case study where the training image is actually identical to the model they want to simulate. This part is very extensive: the three MPS methods are used and different ways of validating the results are compared. The results of this part are according to me not so interesting since in a real case you never have the model you want to simulate but only one or more training images depicting some general geological concepts of the area. The results are also not surprising: iqsim better reproduces the TI which is logical since iqsim uses relatively large patches instead of pixels. In real cases, however, you don't want an exact reproduction of the TI but you want to simulate another area with similar patterns.*'

AC1: We disagree that from a practical point of view the first set of tests are not as interesting since the model we wish to simulate is usually not available. Instead, we consider this case a 'best case scenario' where the TI provides an accurate rendition of the 3D patterns relevant to the given model.
Some of the authors have recently submitted a research paper to HESS, which focuses on uncertainty related to the MPS setup, as well as a presentation of a more practical application, where a 3D geological model from another area is used as TI to simulate a hydrostratigraphic model using SkyTEM data and lithology logs (hess-2017-734; https://doi.org/10.5194/hess-2017-413).

AC2: Regarding the iqsim results we were actually surprised that they performed the best in resembling the cognitive geological model using the Modified Hausddorff distance (MHD) measure. If you look at the individual iqsim realizations (FIGURE 8D) you will quickly realize that the overall placement of the hydrostratigraphic units are not very precise, compared to snesim and DS (FIGURE 8B&C). In fact by looking at the vertical cross-sections the sporadic nature of the upper part of the realizations becomes clear, where the valleys (filled with 'sand & gravel' and 'glacial clay' can be covered by 'hemipelagic clay', which is not possible in the TI. Looking at the borehole distance results, iqsim has the highest average borehole distances for 'sand & gravel' and 'glacial clay' units.

RC2: '*In the last part of their paper, the left half of the existing geological model is used as a TI to simulate the right half of the model. For me, this second part is much more interesting. However, this part is very short: only* *one MPS method is used and different aspects of validation (such as comparison with boreholes) are not shown or discussed. I would like to see an application of the three MPS methods here and a more thorough description and discussion of the results as for the first part where the TI is equal to the result you want to obtain. For clarity and compactness of the paper, I would even propose to only do the full analysis on the second problem where another area is modeled and to remove the part where the TI is identical to the model.*'

AC3: As stated above in **AC1**, *w*e disagree with the statement that the second part is "more" interesting than the first part. Again, we consider the first case a *"pretend case" where* the TI contains the actual 3D patterns of the target model and is therefore a "best case scenario".
Furthermore, focusing on the second case, the half-sim case, a problem occurs since the TI is suddenly cut in half. Cutting the TI in half results in a reduction of the patterns contained in the TI and, as discussed by e.g. Emery and Lantuéjoul (2014), if the size of the patterns contained in the TI are too small we do not properly reproduce the desired patterns. If they become too small the information is simply not available. In this case the valleys are cut in half and only part of the valley structures are present in the half TI. Therefore, although the second half sim case is more interesting from a practical point of view, it is simply not an ideal setup for making such tests.
A recent paper has been submitted to HESS (hess-2017-734; https://doi.org/10.5194/hess-2017-413) where snesim is used to test if we can use 3D hydrostratigraphic models from a different survey area to create 3D hydrostratigraphic models.
The exact same half sim case, with the exact same data, was also presented for iqsim by Hoffimann *et al.* (2017). We will add a reference to Hoffimann *et al.* (2017) in the section describing the half-sim case and include the results by Hoffimann *et al.* (2017) in the discussion of the half-sim case.

RC3: '*Abstract, line 13 + introduction, lines 32-37: I would replace "hydrological" models by "hydrogeological models" or "groundwater models" as "hydrological" models could also refer to surface water modelling, rainfall-runoff modelling or river modelling which do not involve inclusion of geological and/or geophysical data.'*

AC4: Good point, we will change that during the revision.

Response to anonymous referee #2

General comments:
'*This paper provides an exhaustive comparison of three Multiple-Point-Statistics (MPS) methodologies - namely, Single normal equation simulation (snesim), Direct Sampling Simulation (DS) and Image quilting simulation (iqsim) - for the generation of random distributions of hydrofacies on a specific field site. For each methodology, the diverse realizations of hydrostratigraphic categories are obtained on the basis of 51 stochastically-reconstructed resistivity grids, to include the effect of uncertain conditioning (soft) data. The generated hydrostratigraphic models are compared against each other and against the Training Image (TI) (i) by visual inspection, (ii) in terms of the modified Hausdorff distance and (iii) in terms of the distance from borehole (hard) data. The paper is clearly written and the results will have wide application in the con-text of field-scale stochastic facies reconstruction. I recommend the paper for publication in HESS, after that the authors address the questions/comments in the following itemized list.'*

RC1: '*Advantages and disadvantages of each methodology are extensively discussed, and can be summarized as follows:*

> *(i) snesim is the best one in conditioning the simulations with soft data, thanks to the implicit Resistivity Atlas histograms. This methodology provides the best results in borehole distance for 2 out of 3 hydrostratigraphic categories. However, the resulting stochastics models are affected by unrealistic small scale variability, which implies a larger distance from the TI.*
> *(ii) iqsim is the fastest algorithm amongst the three. It provides the smallest distance from the TI and the largest variability between realizations. On the other hand, it suffers from an improper conditioning from soft-data grids, as indicated by poor borehole distance results.*
> *(iii) DS is the most computationally expensive, it suffers from small-scale variability (line 56) and hydrostratigraphic units are not conditioned properly (line 753). It provides intermediate results in terms of all comparison metrics considered.*

So, why did the authors choose DS as the unique methodology in the "Hydrostratigraphic modelling of new surveys", in Sect. 4.3? I would recommend to integrate this section also with the results of the other two methodologies for the simulation of "Area B".'

AC1: The choice of method was not crucial here since the "Hydrostratigraphic modelling of new surveys" was only meant as an example of a practical application of MPS in relation to 3D hydrostratigraphic voxel modelling. Since the DS method is easy to parameterize and easy to setup for running in parallel on a computer cluster it was chosen over using the SGeMS implementation of snesim. This should probably be mentioned in the revised paper.
Regarding integrating the other methods in Sect. 4.3 see author comment 3 and 4 in the response to anonymous referee #1.

RC2: *'The absence of small-scale variability in single realization (iqsim) is regarded as an advantage. But, (1) as discussed in lines 733-741, this reconstructions can be regarded as the most realistic only if the TI is actually reproducing the correct scale of variability; (2) it is the model ensemble, and not the individual random realization, that is supposed to reflect the behavior of the whole system. Small-scale variations effect seem indeed to be reduced when evaluating the mode over the 10 realizations in sect. 4.3. The ensemble modes evaluated over each one of the three sets of 51 simulations analyzed in the first part of the study should be also reported.*'

AC2: The absence of small-scale variability of a single realization seems to be part of the reason for the smaller MHD_{cog}-distances in the iqsim realizations. So in comparison with the TI, which does not contain small-scale variability, the iqsim realizations are the most similar, not realistic. We will revise the text so that it is clear that small-scale variability does not mean less realistic realizations. Furthermore, a new figure presenting the ensemble modes for each of the three algorithms will be considered strongly for the final draft.

RC3: '*It is not explored in this context how the three algorithms behave when generating random simulations with fixed conditioning data. What are the effects of the methods themselves on, e.g., the variability between realizations?*'

AC3: We are not sure what the referee means by "generating random realizations with fixed conditioning data". We assume what is meant is to run realizations with borehole data as hard conditioning data. The usage of hard borehole data for conditioning was not important for the goal of this paper, which was focused on comparing MPS algorithms using an extensive soft SkyTEM data set. However, a recent paper has been submitted to HESS (hess-2017-734), where snesim realizations are conditioned to both soft SkyTEM data and hard borehole data.

RC4: '*line 458: "Here, sand & gravel and glacial clay were categorized into a single category, and hemipelagic clay was used as a background variable". The Modified Hausdorff distance is evaluated on binary images. Did the authors try to evaluate a MHD array separately for each category (similarly to what it is done for AEBD)?*'

AC4: This is a good observation, and should probably be stated more clearly in the revised manuscript why we make an evaluation based on binary images. The reason for this was the computational overhead of computing the Modified Hausdorff Distance for 51 models containing 1,187,823 cells (229x133x39). Even after representing the geometric objects of each realization as outlines only, the computational burden was still too large for computing the MHD for each separate category.

RC5: '*line 726: "The borehole distances of the iqsim realizations revealed exceedingly small hemipelagic clay distances, with average of 0.2 m"; line 730: "(...) the ample near surface hemipelagic clay decreases the hemipelagic clay borehole distance". If the presence of near-surface hemipelagic clay is an artifact of the algorithm (i.e. is not consistent with borehole data), why should it results in a decrease of the borehole distance?*'

AC5: The text will be revised and should instead reflect that if hemipelagic clay is present at the surface then the average MHD increases, and the reason for the low hemipelagic clay distances should be found elsewhere. Instead, in the revised paper, it will be made clear that there is a trade-off relationship between the borehole distances of each of the three lithological categories. In the iqsim case the average distance is low for hemipelagic clay (0.2 m) while increased for the glacial clay (3.5 m) and sand & gravel (5.8 m) categories. The summed distance of the three lithological categories is therefore 9.5 m for iqsim, while the summed distance for DS is 8.5 m and for snesim the distance is 7.5 m.

RC6: '*Figure 2: the figure caption and the references to the figure in the manuscript are not consistent with the letters (a-g) indicating the diverse frames of the picture.*'

AC6: *This is not intended and the figure caption will be edited so it corresponds to the actual figure.*

RC7: '*Eq. 2: Symbols a_i and b_i represent position vectors, but they are written as scalar quantities.*'

AC7: *The symbols will be italic to indicate that they are vectors in the revised version.*

RC8: '*line 536: "where binlog_i is the ith cell in the binary log grid" should be changed into "where binlog_i is the ith ACTIVE cell in the binary log grid".*'

AC8: *Noted, the text will be edited for the final draft.*