Line numbers in blue refer to the edited version

- **RC1**: 'Comment on essd-2021-304', Anonymous Referee #1, 26 Oct 2021 reply

  General Comment:

  The manuscript presents a large dataset that focuses on the geometry of the area of interest and can be useful for training machine learning models in geosciences. The dataset is a step forward in providing the data science community with geoscience related dataset. The authors also recommend several use-cases of the dataset and future works. I have a few minor suggestions/comments to improve the quality of the manuscript. I also feel that the quality of writing can be improved. The models/model history are available on Github as described with a convenient Jupyter Notebook for users.

  Detailed Replies:

  In generating a geologic dataset, users typically have specific parameters of interest. Is there a way for users to conveniently select uncertain parameters to test within the generated dataset?

  The Jupyter notebook allows models to be filtered by event sequence, and with modification could be filtered for any parameter. This point is now made in section 6.5.

  More generally, outside the scope of this manuscript, the pynoddy code provides this option for structural parameters as described in Wellmann et a., 2016, although it does not manage petrophysical variations. We chose to modify the base c language version of Noddy to speed up the calculations. In the future the two codes will probably be merged.

  Line 100: Perhaps the authors should highlight why the publication of a 1-million dataset is needed when users of the Noddy platform can generate up to 3-million models as mentioned by the author in line 100 - in most cases, users typically want to test specific features as opposed to all general features and may have to regenerate the dataset.

  We of course agree that focusing on a subset of structures when the targeted structure is reasonably well characterised is a valid approach, as was taken by Guo et al. 2021, and discussed in section 6.5. Unfortunately, in many parts of the world there is no outcrop available, due to tens to hundreds of metres of cover. In this scenario, it makes sense to start with a broader search for possible 3D models that may match the observed gravity or magnetic response, given their inherent ambiguity. We can imagine a hierarchical approach where a subset of the 1M models is identified as possible causative structures, and then these are accepted or rejected based on the

geologist's prior knowledge, and the accepted models are then used as the basis for a focussed parameter exploration. This point is now made in section 6.5.

Line 148-149: "The likelihood of folds, faults and shear-zones are double the other events as we found that they had a bigger impact of changing the overall 3D geology" - is there a way to illustrate or quantify this?

This is certainly an interesting question related to the impact of different event types on 3D geology, but for this study it was a qualitative observation, and we believe a quantitative investigation is beyond the scope of this paper. This has been clarified in line 155.

Line 151-152: Perhaps highlight how the sampling method (combinatorial versus MC) affects the generated models?

Whilst a combinatorial approach may in theory explore the parameter space more uniformly, the sequence of 5 deformation events is so non-linear that it was reasoned that a pure MC approach would serve our purposes. This point is now made on line 158.

Line 240: Line 56 says that the focus is on six challenges but here it mentions that the authors attempt to address four recognised limitations.

The two missing challenges (Multi-resolution Data & Noise, Incompleteness, and Uncertainty in Data) have been added to line 251.

Line 242-243: Not clear what is meant by "Contrary to the current trend, the work for the generation of a comprehensive suite of geological models did not depend on the appropriate training of a neural network".

This has been rephrased to make it clear that we do not rely on the manual labelling of datasets on line 252.

Line 244-246: Worth mentioning that the problem with GAN is not the amount of samples that can be generated (as the sampling process is fast), the quality of generated samples are limited by the number of training samples used, as well as the stability of GAN in generating realistic samples.

We agree and this point has been added to line 256.

Line 485: Figure 3 is not called anywhere in the manuscript

Fig. 3 is in fact called on line 260.

Minor Comments:

All minor corrections have been applied

Line 30: “applied” -> “application”?

Line 46: best to be consistent with either “data set” or “dataset”

Line 50: be consistent with capitalization

Line 62: “varies” -> “vary”?

Line 132: Extra parenthesis, “python” -> “Python”?

Line 179: “toto” -> “to”

Line 193: “citations” needs to be updated

Line 211: “often?” needs to be updated

Line 306: “started”?

Line 317: “start in”?

Line 358: “in volved”