Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2021-304-RC1, 2021

Comment on essd-2021-304

Anonymous Referee #1

Referee comment on "Into the Noddyverse: a massive data store of 3D geological models for machine learning and inversion applications" by Mark Jessell et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-304-RC1, 2021

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 Comments:

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?

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.

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?

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

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.

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”.

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.

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

Minor Comments:

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”