General comments:

This manuscript tackles a highly topical and important concept of interpretation uncertainty, with a particular focus on 3D seismic. An extensive and appropriate data set is presented which comprises 78 seismic interpretations of the Gullfaks field undertaken by undergraduate students. Interpretational differences in horizon and fault stick picks are reported and quantified, with variability attributed primarily to changes in seismic data quality. The authors go on to discuss how these results have implications for deterministic and stochastic geo-modelling and machine learning.

Whilst the results are well presented, there are occurrences where a stronger analytical framework would lead to greater confidence in the results and improve the scientific rigor of the manuscript. I suggest the authors include an additional section to the discussion to further elaborate on the reasons behind the results, prior to discussing the multiple implications. The addition of a conclusion section would also give readers a brief overview of further work and the manuscripts key points, along with providing a clear take home message. Overall, I suggest that the paper would be of great interest to the readership of Solid Earth and should be accepted if the suggested revisions are implemented. I very much look forward to seeing the revised version of this manuscript and believe this contribution has the potential to become an important contribution to field of 3D seismic interpretation..

Major comments

1. ***Framework for results and use of mean/SD/median:***

The authors provide an extensive and interesting suite of student interpretations, however, a primary concern whilst reviewing the manuscript was the framework and analysis used to describe the results. My concerns can be split into two: (a) the statistical analyse the data and (b) the language and framework used when describing these results.

(a) Throughout the manuscript the authors use a combination of mean, standard deviation, median and IQR, however it is not always clear how the results are distributed. Standard deviation is often used as a measure of uncertainty, however, I have two potential issues with this. The first being that in some cases the results don't appear to be normally distributed (e.g. interpretation density in Fig 9). In the case of a skewed distribution standard deviation should be avoided, and I advise the use of Inter Quartile Range (IQR) instead. It would also be useful to see what the maximum uncertainty, and not just the IQR or standard deviation, with minimum/maximum values reported. My second issues with the use of standard deviation is that a specific standard deviation value could represent either a high or low value of uncertainty depending on the mean at that point. For example, in your fault throw data for F3 (Fig 7b) the standard deviation remains ~30 in the north of the profile. The median throws in this section however ranges from ~50 to ~120, meaning the level on uncertainty is considerably higher at a smaller throw. I suggest to give the reader more confidence, and to better understand the risk in different areas that the following should be considered when analysing the results:

    i. The shape of the distribution? Is the distribution the same along the whole profile?

    ii. If the distribution is not normal, then consider using median and IQR to describe the results.

    iii. The normalisation of results would aid the comparison of uncertainty; this should either be through the use of coefficient of variance, or in the case of non-

normally distributed data then a quartile-based coefficient of variance (IQR/median).

(b) Throughout the manuscript there are several instances where the authors use subjective language when describing the results (e.g. 'significantly reduced', 'observe large', 'roughly correspond'). The description of the results would be improved through the quantification, or further description, of what the authors mean. It would also give readers more confidence in the results. In addition to this there are occurrences where statements such as 'a few students' are used (e.g. Page 7 Line 7). In these cases, the authors should be explicit in what a few means '7 students'.

2. *Discussion section 'key findings'*

This paper provides an impressive analysis of a 1st pass data set aiming to quantify uncertainty in 3-D seismic interpretation, however, I felt the authors needed further elaboration on the reasons behind the presented results. The manuscript would be improved through reducing the key findings to a set of points (potentially even using bullet points), and adding a section which investigates the factors behind the presented uncertainty. From your results this appears to be split into two 'themes':

(a) Human factors: Several sections allude to how students build a mental model during interpretation (e.g. taking information from outside areas of degraded seismic image quality to inform decisions). It is somewhat lacking that the current literature in this area is missing, and that this aspect remains unexplored. Do those who use outside trends to inform areas of poor coverage end up with a better interpretation? Another important comment you raise can be found on Page 13 line 5-7 where you allude to the order students undertook the analysis. This is another important aspect to consider and should be explored further, possibly referencing other work which shows people can vary interpretations through time (e.g. Scheiber et al., 2015 for lineament extraction). The role of human factors on the collection of 3D seismic data should be further explored.

(b) Technical factors: These are highlighted strongly in that uncertainty is higher in areas of lower seismic resolution. Seismic resolution will always be lower surrounding fault, due to the increased amount minor structures and local deformation, and such we can expect uncertainty to remain high in such areas. We know from other aspects of fault science that 'intersection zones' or larger offset faults tend to have a wider zone of damage, and hence zone of reduced seismic image quality. Can we use some of this information to aid the assignment of risk in these areas?

I was often left asking 'why is this the case?' and the answers weren't forthcoming in the discussion. Although I have provided examples of what I feel should be expanded upon, there where sections which could also be expanded and linked to published literature. If this section is added, some of the implication sections could be slightly scaled back in particular section 4.4 which I feel has the least direct link to your results.

3. Conclusions:

I felt this manuscript really lacked was a clear finale. The authors present an extensive set of results, which have clear implications to the interpretation of 3D seismic, however, in my opinion fail to leave the reader with a clear take home message. This point links to the previous regarding the discussion and believe the discussion should be slightly restructured as above and a set of conclusions included which pulls together the findings, highlights the clear importance of these results (including beyond 3D seismic interpretation e.g. modelling from other sources) and raise future research directions. This would tie an important contribution together, and provide readers with a clear take home message.

4. ***Figure quality/readability:***

Some of this may be due to the uploaded PDF, however, I found several figures difficult to read and often containing areas where text was too small. Some examples include line-weights of sub-sections, text size of longitude/latitude and labels within panels and occasionally the chosen colour scheme used was difficult to read either on the screen or when printed off. Detailed points are raised in specific comments.

Specific comments [page (line)]

Abstract: Likely to need minor edits following the suggested edits to the manuscript.

1 (17-21): The introductory paragraph of the MS should be expanded to further define uncertainty. Conceptual uncertainty is first stated to be important in the 2nd paragraph, however, non-specialised readers would benefit from an explicit introduction to the different types of uncertainty (e.g. Bond, 2015; Tennert, 2007). And potentially how this effects the mental model of the interpreter.

2 (31-33): I suggest you make it clear that the study focuses on an interpretation boundary of a student exercise here, it currently sounds like it focuses on an area of a larger dataset.

3 (Fig1): (a) Colour and line weight for section line and interpretation box is unclear, both in colour and in B&W. I advise a change in colour and that the line weight is increased. The text size in the insert to this panel is far too small, as is the longitude and latitude numbers. The addition of a scale bar to this panel would also aid the reader. (b) A scale should be added to this panel. (c) The formation names are poor quality in the uploaded PDF, and also slightly on the small side.

3 (2-3): Many questions come to mind with respect to the level of experience of the interpreters and in part the limitations of your dataset, which includes undergraduate students only. Some of these include: Did everyone have the same level of training? What was the 'specialisms' in the sample set (i.e. how much seismic interpretation, structural geology, stratigraphy etc. was covered and was this equal in the students)? Also how long was spent by each student (If you have this data it would be interesting to see if those who spent more time interpreted differently to those who did not)? How comfortable where the students with using petrel & integrating well and seismic data?

3 (5): How much assistance was given in this? What was the variability in the interpreted horizon when assistance was given and how does this compare to the Top Ness. Can the difference between the Top Cretaceous and Base Cretaceous/Top Ness horizons show the effect of training in reducing uncertainty?

Also if there is little variability in the Top Cretaceous, due to the supervision, will this not effectively 'pin' one end of the fault sticks to a lower range of displacements, effetely adding to the increase in U/C with depth attributed to a degradation of seismic image quality (I agree image quality decreasing with depth will also be a factor).

3 (7-8): Was there any difference in interpretation from students who used these different methods? How often was seeded tracking or manual interpretation used?

4 (Fig2): Increase the text size on the axis for clarity.

4 (2-3): Suggest the text about 90 interpretations be removed as is only mentioned here, and does not seem required.

4 (10): This is an impressive data set, however, I would be interested to see how this is spread between the students. I suspect, and you allude to on page 4 line 15, that the number of fault sticks interpreted varies extensively between students, and that this is an important aspect of uncertainty. This could also then be further analysed to see if there is a correlation between number of fault sticks and level of uncertainty.

4 (15): How is interpretation density defined?

5 (Fig3): I wonder how Fault 1 and Fault 2 are defined in the northern interpretation bin once they are merged.

5 (13): I worry that this is effected not only on the placement, but also on how many fault-sticks each student included. In areas of relatively certain offsets, which will likely be increased by the image quality, I would imagine more sticks will be chosen, thus increasing the apparent 'certainty' of the result.

6 (Fig4): I find the addition of the mean fault plane & k-values from Fossen and Hesthammer (1998) confusing as is, however, it is an important point which you make on Pg 11 ln 31-32. It would be made clearer to the reader if this mismatch was raised in the results, and later discussed in the 'Key findings'. A reminder that stereonet plots go from N to S actually on the figure and not just in the figure caption would also be helpful in this figure.

6 (6-7): I am struggling to pull three clusters out of the stereonet data presented in fig 4a, and instead can only see two. I agree the data should be split into three due to the sinusoidal shape based in the geographic location, however, this information is instead better portrayed in Fig 3a. I advise you reword accordingly.

7 (Fig5): I would like to know the skewness of the distributions, particularly if this changes down dip, this will impact how valid the use of standard deviation is (See major comment 1). I also wonder which fault show the most variability with depth and why. Comparing using either a coefficient of variance (if distributions are normal) or quartile based coefficient of variance could pull out more trends between the faults.

Also although standard deviation increases with depth, how well the data fits the regression line seems to decrease, particularly for F1. For F2 and F3, and to some extent F1 there seems specific horizons which show increased/decreased spread which is not in agreement with the linear regression. Is there an underlying control here? (e.g. stratigraphic layer with good/poor seismic response?).

Visually I would consider changing the 'picks above BCU…' from light grey as it is difficult to see, the regression lines for F2 and F3 are also unclear when viewed on the screen (fine when printed).

7 (7): How many students did this? This is a source of error/uncertainty and I feel it should not be dismissed. What training/geological information was provided to the students and from this should they have factored in the 'geological unreasonableness' of the interpretations?

7 (12): I question why probability is quoted here, you have 78 interpretations, so feel that the numbers represent the total number of students who interpreted that network.

7(14): I feel this needs to be linked back to interpretation and not to 'probable'. Probable suggests that if 100 random people where to be selected then X% would choose option Y, which I think is misleading as there are more human factors involved here. I also feel it is prudent to describe in the MS the level of exposure students has with 'complex' fault topologies.

8 (Fig. 6): In part (a) I would advise that the y-axis is changed to # of students and not a percentage (see comment #). In part (b) I wonder how statistically different A & C are in the students data? Is there a distinct gap? (as topologically they are the same, and geometrically similar).

8 (6): How do you define 'relatively constant' uncertainty? How is it measured? See major comment 1.

9 (Fig 7): This figure makes a very important point, that uncertainty can vary spatially, however, a number of questions are raised in how the results are presented. My main concern is the use of median and standard deviation (Again see Major comment 1). Why is median used? If it is because the distribution is skewed, which I suspect it is, then it is not statistically robust to use standard deviation. I would also like to see the min and max values here (aka what is the maximum risk in this data set?). I suggest redrafting to either show standard deviation surrounding the mean, with min and max values displayed, or to show the IQR around the median again with min/max values. I prefer the second method and suspect similar trends would be observed. Visually I would consider increasing the text size of the annotations.

Is standard deviation in any way related to throw? A +/- 30 meters on a 120 m offset fault is much better (25%) than on a 50 m offset fault (60%) Is quoting exact values the best way to compare uncertainty?

10 (9): How many students interpreted the fault further to the East?

10 (11) to 11 (7): This section suffers from a lack of statistical analysis, a framework to describe these results would increase the rigor of this section. The data shows some very important trends, probably the most important point of the manuscript, and with a more robust statistical analysis the reader would have more confidence in the results and following discussion.

11 (10-18): You open this paragraph with a statement that you show that u/c is correlated to seismic reflector strength, then backtrack on line 13 to discuss human factors. I would suggest that either this paragraph is split and both sections elaborated, or that the topic sentence incorporates both concepts. See Major comment **2**

11 (26): How strong is the Top Ness horizon? Does this effect how well it is interpreted?

11(31-32): How did Fossen and Hesthammer (1998) get their pole? What was there scale of observation (i.e. did they have the data to extrapolate the sinusoidal shape)? The work on this should be included in this part of the discussion.

12 (Fig 9): How do you define interpretation density; units should be added if applicable? Visually this figure could do with a general text size increase, with many areas of text being too small. I would also suggest a change of colour for the boxes in part (1).

13 (5-7): This is a potentially important point and raises a very important question 'what order did students interpret the cube?' If students are spending more time on a certain area, where data is of

better quality then there are more factors to consider in why your results are different. Also does the style of interpretation change with time? I advise either that the key findings section be reduced to a summary (e.g. set of bullet points) and separate section added to explore the reasons behind the uncertainty, probably split into 'technical' (e.g. image quality) and 'human' (e.g. different mental models) and that appropriate literature be added to this discussion.

14 (27-29): I think it would be unwise to suggest normal distributions, even in areas of good seismic data. I suspect in nearly all cases the distributions will be skewed. Most faults display an asymmetric damage zone, and such will also show an asymmetric signature in seismic, should the flat tail be towards the hanging wall?

15 (3): I found this an underwhelming end to a really neat dataset. Although the implications for machine learning are indeed relevant, I feel the MS is crying out for a conclusion section which ties the findings together and includes the 'next stages' in tackling uncertainty in 3D seismic interpretation.

The section itself also seems somewhat out of the remit of this work, and could conceivably either be reduced or cut to make space for a discussion into the reasons behind the results as suggested previously.

Please find additional minor comments/suggested text edits on the attached MS (many of which are included in the specific comments).