

Author Response to RC1 (Billy Andrews)

Major comments responses

[1] 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 analysis of the data and (b) the language and framework used when describing these results.

[1a] 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 issue 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 of uncertainty is considerably higher at a smaller throw. I suggest giving 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).

AUTHORS COMMENT: Thank you for the suggestions! They enabled us to make the manuscript more robust with regards to the statistical description of our findings! We agree with the reviewer in that the use of IQR instead of standard deviation is a better way to analyse the fault throw uncertainty.

CHANGE IN MANUSCRIPT: We have redrafted Figure 7 (now 6) as boxplots of the fault throw uncertainty data and have adapted our description of the findings accordingly.

[1b] 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'.

AUTHORS COMMENT: Thank you for pointing this out! We have adapted the manuscript to remove the use of subjective language where possible. We have improved the statistical rigor in our description of results.

CHANGE IN MANUSCRIPT: A multitude of changes throughout the manuscript. Please see the revised manuscript or the change-tracked manuscript.

[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':

AUTHORS COMMENT: We have structured our discussion into four parts to first present to the reader the key findings of our study, and then to address the implications on what we believe to be the most important applications of seismic interpretation. We have added a summary section though, to provide the reader with a clear take-home message following our discussion.

CHANGE IN MANUSCRIPT: We have done various changes throughout our discussion, please view the revised or change-tracked manuscript.

[2a] 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.

AUTHORS COMMENT: We have no real way of telling which students used outside trends to inform their interpretation in areas of poor coverage. We also have no information about the exact order students interpreted specific faults or fault blocks, as this would have required capturing and analysing their screen throughout the entire interpretation process. We therefore try to abstain from speculation about psychological effects for which we have no evidence for the dataset. We think these factors are vitally important to understanding all sources of interpretation uncertainty, but our manuscript is concerned with quantifying the effect size of this uncertainty and connect it to seismic data quality.

CHANGE IN MANUSCRIPT: -

[2b] 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.

AUTHORS COMMENT: The assignment of risk or uncertainty based on fault offset is not straight forward in our opinion. Our study results have shown, that the seismic data quality around e.g. the

southern part of Fault 3 is low in combination with low throw. Seismic data quality is affected by so many aspects from the sedimentology and structural geology of the subsurface imaged to the imaging itself and the processing. Our study attempts to bypass exactly this to allow for estimation of uncertainties by using seismic data quality itself as a proxy for it (Fig. 8). Thus we believe the in-depth discussion of the multitude of impacting factors on seismic data quality out of scope for this manuscript, especially because it is very challenging research to identify the amount of impact every of the aforementioned aspects have.

CHANGE IN MANUSCRIPT: We have added references regarding the impact of fault zones on seismic data quality for readers to follow.

[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.

AUTHORS COMMENT: We have added a summary/conclusion section to the end of the paper to provide the readers with a clear take home message

CHANGE IN MANUSCRIPT: Please see the revised manuscript for the Summary section.

[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 color scheme used was difficult to read either on the screen or when printed off. Detailed points are raised in specific comments.

AUTHORS COMMENT: Thank you for pointing this out. We have adapted the figures to improve text sizes and legibility of colours and line weights where applicable.

CHANGE IN MANUSCRIPT: Figure changes throughout the manuscript. Please see the revised manuscript.

Detailed comments responses

1 - Introduction

[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; Tannert et al., 2007) and potentially how this effects the mental model of the interpreter.

AUTHORS COMMENT: We agree that a more specific description of the subjective uncertainty of interpretation and how it contrasts with objective measurement uncertainty would increase clarity of the manuscript.

CHANGE IN MANUSCRIPT: Our work is thus concerned with quantifying the scope of uncertainties in seismic interpretation, which represents inevitably biased, human judgment under uncertainty

(Tversky and Kahneman, 1974). This "subjective" uncertainty is in contrast to more "objective" uncertainty (Tannert et al., 2007; Bond, 2015) related to the geophysical acquisition of the data itself.

2.1 – Gullfaks geology and seismic data

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 data set.

AUTHORS COMMENT: We agree, and improved clarity of what area is part of the study both in text and in Figure 1.

CHANGE IN MANUSCRIPT: Our study focuses on the structurally simpler western part of the domino system, where we investigate the uncertainty of the three faults F1-F3 and the fault blocks A-E depicted in Figure 1b.

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.

AUTHORS COMMENT: We have adapted the color and line weights of annotations in Figure 1a to improve legibility.

CHANGE IN MANUSCRIPT: Figure changes.

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 were the students with using Petrel & integrating well and seismic data?

AUTHORS COMMENT: The students all went through the undergraduate studies at Aberdeen, and this was their introduction course to 3-D seismic interpretation using Petrel. We do not have access to the time spent by each student, as they did significant amounts of the interpretation outside formal class times. We agree that this would be an interesting variable to analyse though! We have clarified the student's level of knowledge in seismic interpretation and Petrel.

CHANGE IN MANUSCRIPT: While the students had prior training in structural geology and interpretation of 2-D seismic data, this was the student's first hands-on course in 3-D seismic interpretation using the Petrel software as part of their undergraduate program.

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, effectively 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).

AUTHORS COMMENT: Comparing the difference in interpretations between Top Cretaceous (TC) and both the Base Cretaceous Unconformity (BCU) and Top Ness horizon (TN) to assess the effect of training or supervision on interpretation uncertainty is unlikely to give clear information: The TC horizon is very easy to pick due to its lack of significant deformation and we believe it would not provide a valid comparison with the strongly deformed TN. The faults investigated do not reach above the BCU and thus do not affect the TC.

CHANGE IN MANUSCRIPT: -

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?

AUTHORS COMMENT: We have no information about the specific interpretation tools used by students at any given time. It would make an interesting study though to assess bias introduced by software tools!

CHANGE IN MANUSCRIPT: -

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

AUTHORS COMMENT: We have increased the text size for axis labels / ticks.

CHANGE IN MANUSCRIPT: Changes in Figure 2.

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

AUTHORS COMMENT: While the information is not further discussed in the manuscript, we believe it relevant to mention any (subjective) prior filtering we have applied to the dataset.

CHANGE IN MANUSCRIPT: -

2.3 – Data analysis

(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.

AUTHORS COMMENT: Thank you for pointing this out! We conducted a Bayesian estimation of the differences for the overall standard deviation of fault stick placement between the students falling into the categories of above and below 50% of fault stick interpretation frequency. We have added this to the results section.

CHANGE IN MANUSCRIPT: Analysis of the effect of fault stick interpretation frequency between the students with below-median and above-median fault stick interpretation frequency on overall fault standard deviation was analysed using Bayesian estimation (Kruschke, 2013). We observed a difference in mean standard deviation of \$35.8~m\$, \$20.2~m\$ and \$81.5~m\$ for Fault 1, 2 and 3 respectively, with probability of differences being larger than zero being \$99.3\%\$, \$87.4\%\$ and \$99.9\%\$, making the differences for Fault 1 and 3 statistically credible.

4 (15): How is interpretation density defined?

AUTHORS COMMENT: We have removed the possibly confusing wording of interpretation density (number of interpretations per volume) and just refer to it as frequency.

CHANGE IN MANUSCRIPT: In the following analysis we present 2-D and slices of 3-D histograms of fault interpretations, showing interpretation frequency across the domain.

3 - Results

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

AUTHORS COMMENT: This is dependent on the student interpretation. If they interpreted Fault 1 as continuing or Fault 2, respectively.

CHANGE IN MANUSCRIPT: -

5 (13): I worry that this is affected 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.

AUTHORS COMMENT: Indeed, as this is a basic histogram of the fault sticks, high frequency interpretations can affect the plot. But these differences in frequency are along strike and should have a negligible effect on the fault placement uncertainty patterns observed orthogonal to strike.

CHANGE IN MANUSCRIPT: -

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.

AUTHORS COMMENT: We have added an explanation of the Fossen & Hesthammer (1998) to the main text in the results section to improve clarity. We have also merged Figure 3 and 4 into 3, which will enable the reader to identify more easily which Stereonet plot belongs into which bin.

CHANGE IN MANUSCRIPT: We have added Bingham mean poles from \citet{fossen_structural_1998} for all three faults in the plot (light blue) for comparison.

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.

AUTHORS COMMENT: The clusters correspond to the three bins. You can see the clusters when separated into the three separate Stereonet plots. This should be more clear to the reader now that Figure 3 and 4 were merged. We have updated the figure reference accordingly.

CHANGE IN MANUSCRIPT: Fault 1 shows three distinct clusters of orientations (Fig. 3A, a-d), [...]

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).

AUTHORS COMMENT: Median skewness of the distributions show approximate symmetry to moderate skewness of the distributions, but with significant standard deviation (see Table 1). We concluded from that, that the use of standard deviation is adequate for the purpose of this paper. We use here a basic categorization of -0.5 to 0.5 as approximately symmetric and ± 0.5 to ± 1.0 as moderately skewed. Overall, the variation of skewness with depth is quite high. We are aware that collapsing all the information into a single scatterplot is a significant simplification of the data, but we think this to be adequate for the first-look scope of the paper.

The quality of the seismic response of horizons is a major controlling factor of interpretation uncertainty and will certainly have a strong impact on uncertainty. As we have shown in Figure 9, strong bounding seismic reflectors significantly reduce interpretation uncertainty of faults. In this plot we collapse the entire fault uncertainty along the y-axis of the domain, thus averaging across a large area which makes the correlation with specific physical features challenging, and we did not pursue for this manuscript.

Table 1: Median skewness for fault position with depth.

	Fault 1	Fault 2a	Fault 2b	Fault 3
Median Skewness	-0.06 \pm 0.36	0.25 \pm 0.36	-0.20 \pm 0.33	0.21 \pm 0.27

CHANGE IN MANUSCRIPT: Changed grey Figure 5 text to black for better legibility.

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?

AUTHORS COMMENT: We did not count the number of interpretations that do so, but removed any fault stick point above the mean BCU from the analysis automatically. But we believe that most students simply just did not terminate their fault interpretations correctly, as no geological model was built afterwards. We do not believe that most students who did this actually interpreted what they thought were meaningful faults above the BCU conceptually, but rather due to lack of attention to this detail, as this was not a focus of their coursework.

CHANGE IN MANUSCRIPT: -

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.

AUTHORS COMMENT: Used student counts for clarity.

CHANGE IN MANUSCRIPT: Five modes of FN topology make up the bulk of fault network topologies, while others were only interpreted by 3 or less students respectively.

7(14): I feel this needs to be linked back to interpretation and not to 'probable'. Probable suggests that if 100 random people were 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.

AUTHORS COMMENT: Changed probable to frequent for clarity.

CHANGE IN MANUSCRIPT: Note that the most frequent FN (Fig. 6b, A) is different from the reference expert FN interpretation ...

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).

AUTHORS COMMENT: Added student numbers to the probability mass plot for clarity. We also enhanced the clarity of the related schematic fault networks.

CHANGE IN MANUSCRIPT: Figure 6 changed.

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

AUTHORS COMMENT: We are referring to no apparent changes in the uncertainty pattern of the fault throw along strike. We clarified that in the manuscript.

CHANGE IN MANUSCRIPT: Fault throw uncertainty shows no apparent pattern change along the strike (around $\pm 30\text{m}$), while rising sharply to about $\pm 75\text{m}$ at the southern edge of the seismic cube.

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?

AUTHORS COMMENT: Thank you for pointing this out! We have adapted the plots to boxplots to show median fault throw with IQR, minima and maxima, as well as outliers to improve the robustness of our analysis. Overall the same trends visualize, but the difference in fault throw uncertainty at the merge of Fault 1 and 2 are much more visible. We have modified the results description accordingly.

We have increased font sizes in the figures.

CHANGE IN MANUSCRIPT: Results of the fault throw analysis are plotted in Fig. \ref{fig:07}. The boxplots show median fault throw with the associated interquartile range (IQR), extrema and outlier values along fault strike direction. The throw profile of Fault 1 (Fig. \ref{fig:07}a) shows a distinct sinuous shape spatially associated with its interaction with Fault 2. This shows one bin with high median fault throw of approx. \$180~\text{m}\$ and high fault throw uncertainty before strongly decreasing in fault throw values down to a median of about \$40~\text{m}\$ and one of the lowest IQR along the fault. Median fault throw then rises steadily towards the South while increasing in uncertainty. Notice the increase of uncertainty at both ends of the dataset, with increasing median throw in the South and decreasing in the North. The throw profile of Fault 3 (Fig. \ref{fig:07}b) shows two distinct levels of throw: In the Northern part of the fault median throw values are high at around \$90\$ to \$105~\text{m}\$, associated with comparatively lower uncertainty than in the South. Towards the South, median fault throw decreases down to about \$40~\text{m}\$ while the IQR values increase, being largest at the Southern edge of the dataset.

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

AUTHORS COMMENT: Three students interpreted Fault 3 further to the East.

CHANGE IN MANUSCRIPT: -

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.

AUTHORS COMMENT: We have increased the rigor in our statistical description of the results to increase clarity. Thank you for pointing this out!

CHANGE IN MANUSCRIPT: Several changes throughout the paragraph – please see Section 3.3 in the revised manuscript or the change-tracked manuscript.

4 – 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?

AUTHORS COMMENT: We do not see this as much as backtracking from the first statement, but rather we are addressing the complexity involved: Fault interpretation uncertainty correlates with seismic reflector strength, as our study results show, but it is dependent on many more variables (multiple correlation).

CHANGE IN MANUSCRIPT:

13(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.

AUTHORS COMMENT: Fossen and Hesthammer (1998) calculated the pole using Bingham analysis of data with some significant spread. Their study was conducted on the basis of “reprocessed 3D seismic (ST8511)”.

CHANGE IN MANUSCRIPT: We have added information on the origin of the data to the manuscript.

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).

AUTHORS COMMENT: The term density refers to probability density of the interpretations. This just means that the integral of the histogram sums up to 1 and is a dimensionless number. This normalization makes the histograms comparable despite fluctuations in fault stick counts across the cube.

CHANGE IN MANUSCRIPT: We have increased text sizes and modified colours in Figure 9 to improve legibility.

15 (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.

AUTHORS COMMENT: Indeed! For such an analysis the entire interpretation process would need to be captured (e.g. screen capturing) and analysed, which was not an available option for this study. We have therefore no information about the temporal evolution of student’s interpretation.

CHANGE IN MANUSCRIPT: -

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?

AUTHORS COMMENT: The skewness of a Normal distribution can be an important factor for the detailed modeling of structures and should be considered wherever possible. But to our knowledge no detailed study has ever been carried out in trying to identify geological controls on skewness and kurtosis for probability distributions describing faults (or other geological structures). Stochastic geomodeling also needs to achieve a difficult trade-off between the level of detail and computational feasibility, as the correct sampling from the joint distribution constructed from complex parameter distributions (e.g. fat-tailed distributions) becomes increasingly challenging given today’s computational constraints (e.g. Betancourt 2017, 2018). Coupling knowledge about asymmetric damage zones with interpretation experiments would be a very interesting future research topic though!

CHANGE IN MANUSCRIPT: -

15 (3): I found this an underwhelming end to a really neat data set. 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.

AUTHORS COMMENT: We have added a conclusion section to the manuscript.

CHANGE IN MANUSCRIPT: Please see the revised manuscript for the added conclusions section.

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

AUTHORS COMMENT: We have considered and incorporated numerous minor comments and text edits from the attached commented MS.

CHANGE IN MANUSCRIPT: Please see the revised manuscript or change-tracked manuscript.

References

Betancourt, M. (2017). A Conceptual Introduction to Hamiltonian Monte Carlo. ArXiv:1701.02434 [Stat]. Retrieved from <http://arxiv.org/abs/1701.02434>

Betancourt, M. (2018). The Convergence of Markov Chain Monte Carlo Methods: From the Metropolis Method to Hamiltonian Monte Carlo. *Annalen Der Physik*, 1700214. <https://doi.org/10.1002/andp.201700214>

Bond, C. E., Gibbs, A. D., Shipton, Z. K., & Jones, S. (2007). What do you think this is? "Conceptual uncertainty" in geoscience interpretation. *GSA Today*, 17(11), 4. <https://doi.org/10.1130/GSAT01711A.1>

Fossen, H., & Hesthammer, J. (1998). Structural geology of the Gullfaks Field, northern North Sea. Geological Society, London, Special Publications, 127(1), 231–261. <https://doi.org/10.1144/GSL.SP.1998.127.01.16>

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2), 573–603. <https://doi.org/10.1037/a0029146>

Lindsay, D. R. (2011). *Scientific writing*. CSIRO publishing.

Tannert, C., Elvers, H.-D., & Jandrig, B. (2007). The ethics of uncertainty: In the light of possible dangers, research becomes a moral duty. *EMBO Reports*, 8(10), 892–896. <https://doi.org/10.1038/sj.embor.7401072>

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>