Response to reviewers' comments on "Development of a Seismic Loss Prediction Model for Residential Buildings using Machine Learning – Christchurch, New Zealand"

Reply to Referee #2

	Referee #2 Comments	Responses
1	Line 317 - The authors, referring to Figure 12a, mention that the model underpredicted 3.4% of the overcap claims. However, 3.4% is the percentage with respect total claims, and not just overcap claims. A more accurate representation will be - the model underpredicted 67% ((47+48)/(47+48+46)) of the overcap claims. This applies to similar conclusions in later sections as well, e.g., line 356.	Thanks the remark. The numbers have been changed to reflect the model performance on the overcap claims only. The following instances have been amended: Line 317: "However, it underpredicted 3.4% of the overcap claims" was changed to "However, it underpredicted **67**% of the overcap claim" Line 318: "14% of the buildings for which a 'medium' claim was lodged were predicted as 'low' and 17% of the 'low' instances were assigned the 'medium' category." was changed to "**40.1**% of the buildings for which a 'medium' claim was lodged were predicted as 'low' and **28.1**% of the 'low' instances were assigned the 'medium' category" Line 321: "13.7% of the instances in the validation set were properly assigned to the overcap category." was changed to "56.5% of the 'overcap' instances were properly assigned to the overcap category" Line 323: "The performance in the 'medium' category was also satisfactory with 18.6% of the instances correctly predicted." was changed to "The performance in the 'medium' category was also satisfactory with 41.5% of the instances correctly predicted." Line 325: "Despite the optimisation of the model on recall, 8.9% and 1.6% of the overcap claims were wrongly assigned to the 'medium' and 'low' category respectively" was changed to "Despite the optimisation of the model on recall, **36.9**% and **6.6**% of the overcap claims were wrongly assigned to the 'medium' and 'low' category respectively" Line 362 (formerly line 356): "On the validation set, the model achieved 0.59 recall on the overcap category with only 5.9% of the overcap instances underpredicted." Was changed to "On the validation set, the model achieved 0.59 recall on the overcap category with **41**% of the overcap instances underpredicted." Line 370 (formerly line 363): "Model 3 only underpredicted 2.5% of the overcap claims. It is satisfactory to see that among those instances only 0.46% having overcap losses were classified as low." was changed to "Model 3 underpredicted 70.6% of the overcap instances with 13.1% of the overcap claims classified as low. The model had difficulties differentiating between the categories medium

		and low. 34.1% of the medium claims were underpredicted as ‘low’.”
2	The authors have presented a compelling analysis in figure 12, but did not include conclusions from figures 12b and 12c in the text. It would be helpful for the reader if the authors commented on those figures, the differences in model performance when tested on different earthquakes, and consequently, any conclusions that can be drawn about model generalization.	Thanks for the comment. A new paragraph has been added at the end of section 8. “Figure 12b and Figure 12c help to understand how each model, trained on 4 September 2010 and 22 February 2011 data respectively, performed when applied to another event. Figure 12b shows that the recall for the ‘overcap’ category of the model trained on 4 September 2010 applied to 22 February 2011 reached 0.24. For the model trained on 22 February 2011 data applied to the 4 September 2010 event, the recall was limited to 0.07 for the ‘overcap’ category with only 7.4% of the ‘overcap’ claims being correctly assigned to the ‘overcap’ category. This shows that besides assessing the performance of a model on a validation set coming from the same earthquake as the training set, it is important to evaluate any ML model on a different earthquake event before making any generalisation.”
3	The authors have consistently used recall as the evaluation metric, but changed it to accuracy in section 13. It would be helpful to understand why the metric was changed, or provide a confusion matrix, as in other sections, comparing their model with the RiskScape v1.0.3 software.	Thanks for the suggestion. Table 4 has now been replaced by Figure 15 showing the confusion matrices for the RiskScape v1.0.3 software. Section 13 has reformulated to clarify that the main evaluation was performed according to the recall in the overcap category. Nevertheless, the overall accuracy is still presented in Fig 15 as it was deemed important to show to the reader the difference in assessing RiskScape predictions on the overall accuracy or recall.