We sincerely thank reviewer 1 for their detailed and constructive feedback, and the time taken to review this manuscript. We have addressed all comments and provide a response below.

Reviewer comments are in black, our responses are in green, and excerpts from the revised manuscript are in blue.

General comments

My main impression from reading this manuscript is the focus on the method. The method is long, and although I understand the complexity of finding a balance between conciseness and thoroughness in describing the development of such a pipeline, some parts are hard to follow, and some aspects remain somehow obscure after multiple reads. Everything is documented below, but two aspects that remain puzzling at this point are:

1. The introduction to the two classifier tasks - which are mentioned early in the method section with reference to a more detailed description that, unless I am mistaken, never really comes

    - For clarity we have added in the following explanation to section 2.3.2:

        2.3.2 Damage classification

        We chose to divide building damage classification into two separate classifications, Classifier 1 distinguishes between 'No damage to minor damage' versus the combined classes of 'Moderate damage' and 'Major damage', while Classifier 2 further differentiates between 'Moderate damage' and 'Major damage'. A hierarchical approach to classification has been found effective when the number of samples is limited or classes are unbalanced (Li et al., 2019b; An et al., 2021). We conducted experiments separately for Classifiers 1 and 2.

2. The generation and use of the orthomosaic, which raises several question (i.e., georeferencing of some of the datasets, whether the training and further predictions are performed on the orthomosaic or individual images).

    - All bounding box labelling, training, and evaluation described in Section 3 is conducted on the individual non-georeferenced images. This was done for two reasons, firstly to preserve the multiple viewing angles that we have in the images; and secondly since Dataset 1 does not contain GPS positioning or altitudes. However, we recognize that for operational purposes spatial information is required, therefore we generated the pipeline which can operate over both georeferenced or non-georeferenced images. To demonstrate its application, we generated the orthomosaic images of the town of Owia, the only location in our testing dataset that has sufficient images with spatial information to do so. We have clarified this in Section 2.3, which now reads:
      Past studies have trained deep learning algorithms on georeferenced images (i.e., each pixel has a geographical location attached) (Gupta and Shah, 2020; Shen et al., 2021; Bouchard et al., 2022) and non-georeferenced images (e.g., Li et al., 2019a; Pi

et al., 2020; Cheng et al., 2021). In this work we labelled the non-georeferenced images and trained models on these. This was done firstly, to preserve the multiple viewing angles that we have of each building with each image counting as a different data point, and secondly, due to the absence of GPS locations on a large portion of the dataset. In an operational context, spatial information must be tied to the assessed damage. Therefore, beyond the creation of distinct models for each task, we designed a comprehensive, fully automated pipeline that integrates models for building localisation and damage classification. Our pipeline contains all of the necessary processing steps to guide images through the separate models enabling them to operate on a georeferenced orthomosaic image (to be generated separately) or on non-georeferenced images. When applied to an orthomosaic image the output from the pipeline is a georeferenced vector dataset that can readily be plotted in a GIS to generate damage maps.

In Section 4 we apply the pipeline to assess building damage in the town of Owia, which is in the north of St Vincent and received 50-90 mm of tephra fall (Figure 1). Owia was selected out of the three possible test set locations (Figure 3) due to its large size and the existence of GPS locations that enabled the generation of a georeferenced orthomosaic image; for this we used Agisoft Metashape software. To compare the assessed building damage with tephra thickness, we used the TephraFits code (Biass et al., 2019) to identify the theoretical maximum accumulation using the isopachs from Cole et al., (2023). This maximum accumulation and the isopachs were interpolated using cubic splines and the surface was exported at a resolution of 10 m to provide a tephra thickness value for each building.

- We did not manually georeference the image datasets that did not have GPS positioning or altitudes. We now specify this in the description of the datasets in Section 2.1:
  Images do not contain GPS positioning or altitudes and were not manually georeferenced.

The second aspect is the balance between methodology and application. Although the case-study serves as a basis for the development of a method, its application in section 4 is less than 20 lines, which felt anticlimatic! Perhaps is there a political context that prevents further analyses as it is often the case following recent eruptions, and I don't expect the authors to remodel the manuscript around a more detailed analysis. However, even without a direct application to la Soufrière volcano, some aspects could be discussed in the context of physical impacts to buildings to widen the very method-oriented message to a broader audience.

- We do not present a full damage assessment for all locations because the bulk of the data that we have were used for training and evaluating the model, and so we considered that applying our model to data that it was trained on was somewhat circular. In the testing dataset there were three locations that were not used for model development: Owia, Richmond and Troumaca (see Figure 3 main text), we used the largest of these (Owia) to run our example application. Nevertheless, we agree that more analysis in the context of damage assessment is a good idea and we have elaborated on the Owia example. We interpolated between the isopachs and extracted

a tephra thickness per building. We have adapted Figure 7 to include plots that show the number of each damage state as a function of tephra thickness bins (see below), and added in the following description of the Owia results:

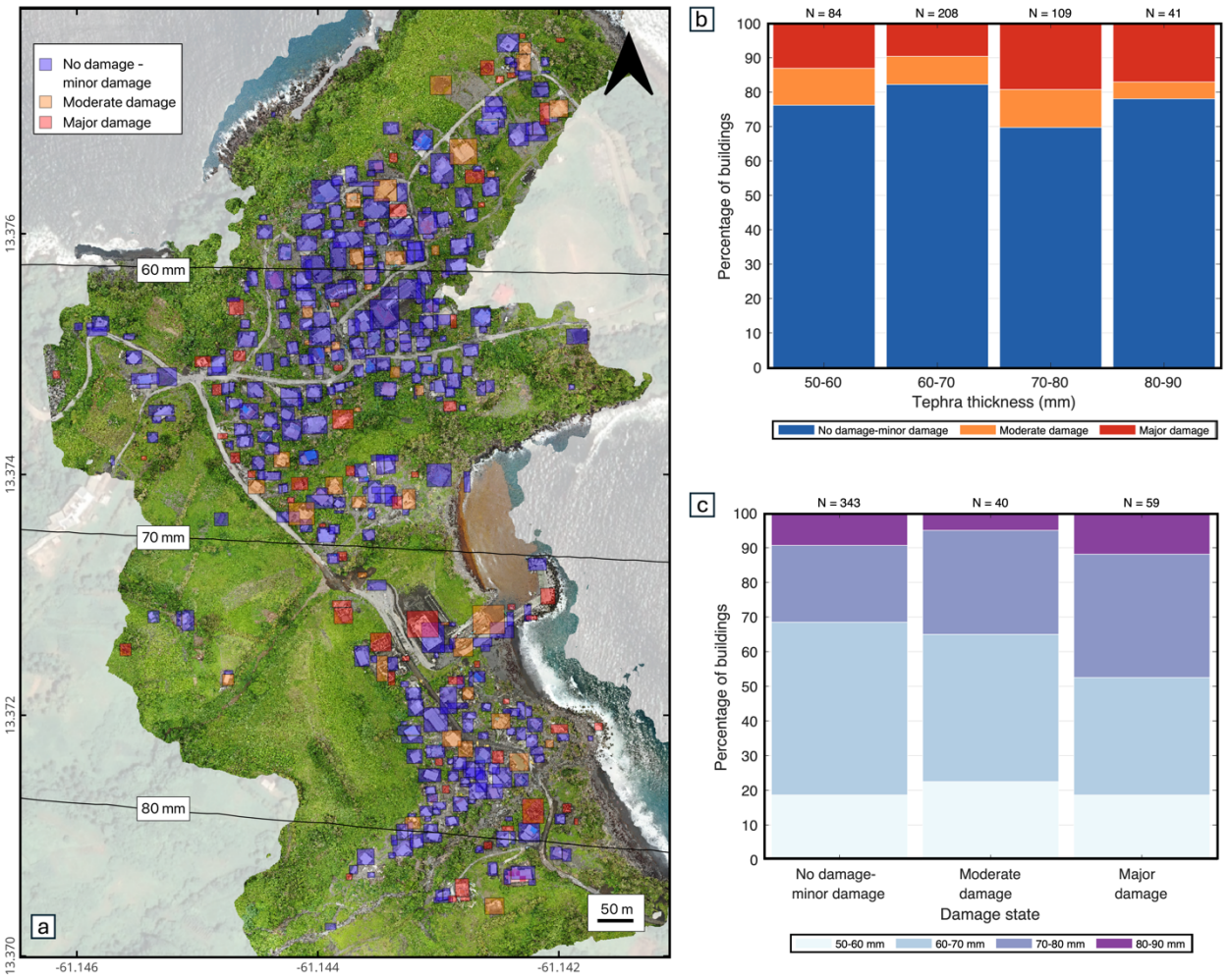
The tephra fall building damage map shown in Figure 7a was produced by overlaying the pipeline output georeferenced vector with the orthomosaic image in QGIS. Our remote damage assessment pipeline identified 442 buildings. Of these, 78% (N = 343) were classified as having No damage to minor damage, 9% (N = 40) as having Moderate damage and 13% (N = 59) as having Major damage. We observed that the two upper tephra fall thickness bins (70-80 mm and 80-90 mm), both had a higher proportion of buildings with Major damage compared to the lower thickness bins (Figure 7b, c), indicating a correlation between tephra fall thickness and building damage though it is not very pronounced. These findings are discussed in Section 5.3.

We have adapted the discussion Section 5.3 to include discussion of our findings for Owia in the context of the damage assessment literature:

Application of our remote damage assessment pipeline to the town of Owia found that 22% of buildings that received tephra accumulation in the range of 50-90 mm experienced Moderate damage or Major damage. Within this range, the relationship between tephra thickness and building damage was not as pronounced as in other studies (Blong, 2003b; Hayes et al., 2019; Jenkins et al., 2024). This may be attributed to the small geographic area and therefore small range of tephra thicknesses considered in our application when compared to other studies. In the damage assessments of Blong, (2003b), Hayes et al., (2019) and Jenkins et al., (2024) buildings received 100-950 mm, 1-600 mm, and 1-175 mm respectively. Spence et al., (1996) assessed building damage over a similarly narrow range of tephra thicknesses to this work (150-200 mm) and found that there was considerable variation in the level of damage despite the majority of buildings having a metal sheet roof. The spacing between the principal roof supports (roof span) was found to be important for the amount of damage observed, with long span buildings experiencing higher levels of damage than short span ones (Spence et al., 1996). There are limited long span buildings in the Owia case study, however additional characteristics such as construction style and material, building layout, age, condition, height, and roof pitch can all affect a buildings ability to withstand tephra loading (Spence et al., 1996; Pomonis et al., 1999; Blong, 2003b; Jenkins et al., 2014). Variation in these characteristics across Owia could be responsible for the observed variation in building damage over the narrow range of thicknesses considered.

If we convert tephra thickness to loading, we can compare the results of our assessment with existing relationships between tephra loading and damage for similar building types. Using a density of 1500 kg/m^2 (Cole et al., 2023) suggests that a loading of at least 75-135 kg/m^2 was applied to buildings for the range of thicknesses considered (50 mm-90 mm). Census data for Owia states that 90 % of buildings have metal sheet roofs (SVG population and housing census, 2012), with the remaining 8% comprised of reinforced concrete roofs and 2% ‘other material’. Given the higher resistance of the 8% of non-metal sheet roof buildings in Owia, we might expect vulnerability models developed for metal sheet roofs to overestimate

damage in the town. Fragility functions developed for Indonesian style buildings with metal sheet roofs (Williams et al., 2020), calculate a 48-80% probability of Owia buildings experiencing damage exceeding Damage State 2, higher than the 22% experiencing Moderate or Major damage in our study. Fragility curves for roof failure (Major damage) of old or poor condition metal sheet roofs (Jenkins et al., 2014), calculate that just over 10% of buildings in Owia would experience sufficient loading for roof collapse, comparable to the 13% observed in our study. These comparisons highlight some of the challenges associated with using vulnerability models developed for different locations. Moreover, they reiterate the need for the collection of building typology and post-event impact data that can be used to increase the amount of empirical data available for vulnerability model development and allow regional vulnerability models to be developed for specific building types.



One component I felt was lacking is the analysis of how the defined damage states fit in a wider damage classification scheme. The only mention to this aspect is found in Section 5.5. However, discussion points raised only seem to focus on the number of classes on the scale, but miss a more systematic comparison with other schemes as well as a critical interpretation of the damages captured by the present

methodology. As a result, the reader is left with a long list of computed parameters to assess the quality of the impact assessment, but with no concrete link to reality. Would it be possible to add:

- Images of a limited number of buildings illustrating each damage?
    - We have added the following figure to show examples of the different damage states.

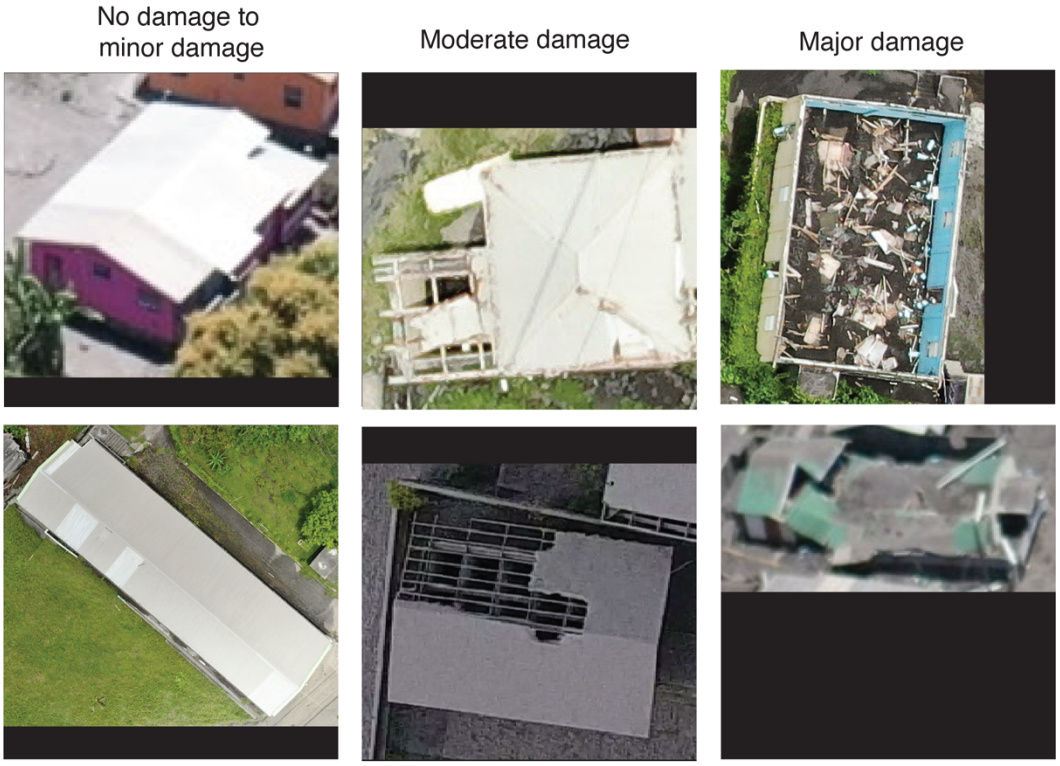


Figure 2. Example of the three damage states used in this work: No damage to minor damage, Moderate damage and, Major damage.

- A description of what these damages capture? → i.e. is the difference between moderate and heavy a structural component? Roof collapse?

    - We have added a new table (Table 2) that shows existing tephra fall damage schemes (Spence et al., 1996; Blong, 2003; Hayes et al., 2019; Jenkins et al., 2024) and have adapted Section 2.2 to include more discussion around how our scheme fits with existing schemes:

Developing and applying a building damage state framework

The first tephra fall building damage state framework was developed after the eruption of Pinatubo, Philippines, 1991 (Spence et al., 1996), and was adapted from the macroseismic intensity scale used to evaluate seismic damage (Karnik et al., 1984). In the adapted framework damage ranges from DS0 – "no damage", through to DS5 – "complete roof collapse and severe damage to the rest of the building". Subsequent tephra fall building damage state frameworks were modified from the

work of Spence et al., (1996) with changes in the wording made to reflect the characteristics of the case study (Table 2). In the damage state descriptions, damage to three critical aspects of a building is described: the roof covering, the roof structure, and the vertical structure (Blong 2003b; Hayes et al. 2019; Jenkins et al., 2024). In our study, most images depict buildings from an at nadir or close to nadir perspective making roof damage more discernible than damage to the vertical structure. Thus, we generated a damage state framework that is based on the proportion of observable damage to the roof, as in the work of Williams et al. (2020). Our final framework, which was developed over several iterations, classifies building damage into three classes: No observable damage to minor damage, Moderate damage, and Major damage (Table 3). Damage states are deliberately generic so that the range of possible damage to the range of different building types can be captured (Blong, 2003a). Our three classes are comparable to DS0-1, DS2, and DS3-5, respectively, of damage scales developed for ground surveys (Table 2). In the frameworks presented in Table 2, DS1 describes light/minor damage or superficial damage to non-structural components. In our framework we included minor damage in the No damage class since the difference between the two can be subtle and not easily discernible through remote assessment. Furthermore, buildings with minor damage are typically habitable and unlikely to require costly repairs; therefore, from a response and recovery perspective, we considered them better grouped with undamaged buildings. Our Moderate damage class requires damage or collapse to up to 50% of the roof area, which closely fits with damage state 2 of Blong, (2003), Hayes et al., (2019) and Jenkins et al., (2024). The ground-based frameworks distinguish damage states 3 through 5 by increasing amounts of damage to the building walls (Table 3). The quantity and severity of impacted walls is not easy to differentiate in the majority of our UAV images, which show buildings from a nadir or close to nadir perspective. Therefore in our framework, we group these states together under 'Major damage'.

- Our damage assessment scheme is provided in Table 3, we have now added to the table a translation between our scheme and existing schemes.
- Regarding the long list of computed parameters we have revised parts of the results (Section 3) to provide higher level oversight of what the numbers mean. Considering this we have also removed one of the figures (confusion matrix-original manuscript figure 6) to simplify the message and to make the writing more direct. Rephrased sections include:
- Section 3.1.1:

  All trained sieve networks achieved macro and class F1 scores that were > 0.973 (Table 5). The sieve networks efficacy at improving building localisation is demonstrated by comparing the results of the best detector pre-sieving (Table 4 row ID 1) with the post-sieving results. Pre-sieving there were a large number of false positive detections, resulting in a precision of 0.588, post-sieving these were reduced and the precision increased to 0.695 (Table 5).
- Section 3.1.2:

Analysing performance variations across different testing datasets can then inform recommendations for future data collection strategies (see Section 6).

- Section 3.1.3:
  To understand if a better model could be achieved with more data available for training, we combined the training and validation data and used this to retrain the best experimental setup for the detector. Evaluation of the retrained model on the test set resulted in an average precision increase from 0.701 to 0.751 for the non-sieved detector, and from 0.668 to 0.728 for the sieved detector, showing that having more data available for training produced a better model (Table 6).

  While the AP is higher for the retrained detector without the sieve, the addition of the sieve network creates a better balance between the precision and recall which is reflected in the higher F1 score (Table 6). For the present application equal importance is given to: 1) making correct predictions about building locations, and 2) identifying as many buildings as possible. Consequently, striking the balance between precision and recall is crucial.

- Section 3.2.1:
  The five experiments with the highest macro F1 score are shown in Table 7, with the full lists provided in Tables S3 and S4 of the supplementary material. For Classifier 1, Macro F1 scores across all 15 experiments ranged from 0.753 to 0.836, while for Classifier 2 scores ranged from 0.776 to 0.810 (Tables 7, S3, S4). Models trained to differentiate between the No damage to minor damage and Damaged classes performed better for the No damage to minor damage class, while those trained to differentiate between Moderate and Major damage performed better for the Major damage class (Table 7).

- Section 3.2.2:
  For Classifier 2, the Moderate damage class is more sensitive to the choice of location and dataset used for the evaluation than the Major damage class (Figure 6). For the different locations the mean F1 score ranged from 0.583-0.974. Similarly to Classifier 1, the location with the lowest mean F1 score is Fitz Hughes, whereas the highest score was produced for Orange Hill. For the different datasets the range for the Moderate damage class is between 0.522-0.746. For the Major damage class F1 scores for the distinct locations are between 0.728-0.933 while for the different datasets the range is between 0.711-0.867.

In any case, I suggest adding - where possible - a couple of bridges that help interpreting the model results beyond the simple application of the method, and broadening findings to the actual literature on impact assessments on buildings (note: the discussion could also be more supported by references!).

- We have updated the discussion to include a section that considers our damage assessment for Owia in the context of the literature on impact assessment, see earlier comment on page 3

- We have revised the discussion to include the relevant references:

  Section 5.1 now reads as follows:

  Through running our building localisation experiments we found that the pre-processing of images before detector training (particularly the block size) significantly influenced detector performance. The block sizes tested were chosen as a trade-off between reducing image size sufficiently to reduce computational cost, and retaining a large enough size such that buildings were not dissected unnecessarily. Given that the optimum block size was the middle size of the range tested, we are confident that this balance was achieved. Cross-validation results demonstrated variability in average precision (AP) for models trained on different locations and imagery datasets (UWI-TV/GOV/SRC) (Section 3.1.2; Figure 5). Deep learning models are known to perform well when the data they are evaluated on have similar characteristics to the data they were trained on, though have more difficulty when working with 'out of distribution' samples (Ben-David et al., 2010). Given the relatively consistent building typology across locations (most buildings observed are detached single storey buildings with either a gable or hip shaped metal sheet roof; a lesser proportion have flat concrete roofs), the differences in AP are likely due to observable variations in UAV altitude, off-nadir angles, tephra thicknesses, and varying training sample sizes.

  The cross-validation AP was notably lower for the London and Fitz Hughes datasets (Section 3.1.2). For the London images (from SRC and GOV datasets) this is likely caused by the smaller apparent size of buildings in these images compared to the other locations, due to the higher UAV altitude. Variations in object size within the training and testing data has been found to affect the performance of deep learning models developed for building localisation, with models often performing better for objects that are the same size as those in the training data (Nath and Benzadan, 2020; Cheng et al., 2021; Bouchard et al., 2022). Fitz Hughes images were all from the UWI-TV image dataset which contributed just 17% to the combined training and validation set used for cross validation. This dataset was collected closer in time to the eruption, therefore as a whole had more tephra on the ground than the SRC and GOV datasets, which affects background colour. Furthermore the UWI-TV dataset viewed buildings mostly from an off-nadir perspective, while the other datasets were predominantly nadir images. The effect of image background colour on localisation performance is expected to be minor, Cheng et al., (2021) found that for the same event localisation AP dropped from 65.6 to 63.3 when their model was tested on images containing buildings surrounded by vegetation compared to buildings with an ocean backdrop. While Bouchard et al., (2022) suggested that models quickly learn to ignore background pixels. On the other hand, differences in off-nadir angle is a widely acknowledged challenge of working with UAV or aerial images (Cotrufo et al., 2018; Nex et al., 2019; Pi et al., 2020). Under representation of the mostly off-nadir UWI-TV images in the training data may have impacted the

model's ability to recognise such instances in the test data. During model development we experimented with different models for the different datasets (UWI-TV, GOV, SRC), but found that models developed on the combined dataset performed better than those developed on the separate datasets and a combined model was the one selected and used for cross validation. Rather than suggesting that variations in off-nadir angle are not important, this finding likely reflects the smaller size of the individual datasets compared to the combined datasets, meaning that less information was available to learn from. The application of sampling approaches like those used for the damage states in the classification model development (over or under sampling) could have been applied to balance the data. However, the SRC dataset is much larger than either of the UWI-TV and GOV sets (Figure 3), therefore we considered that oversampling would introduce significant bias towards the specific examples in the under-represented dataset, whereas through under sampling we would lose a large amount of the data that are available to learn from. Given these factors, we did not use sampling approaches. Future work might consider the application of generative AI algorithms such as generative adversarial networks (GANs) to expand the dataset (e.g., Yi et al. 2018; Yorioka et al., 2020), although more work needs to be done to quantify the diversity in the generated data.

The variability in cross-validation results for the building localisation model likely comes from a combination of the above factors (differences in UAV altitude, off-nadir angles, tephra thickness, and varying training sample sizes), and suggests that there was insufficient information in the training data for our detection models to perform well across the range of characteristics present. This is supported by the increased performance when the best localisation model was retrained on the combined training and validation data. However, further investigation is required to separate the unique effect of each aspect.

Specific comments

Line 84-89: I suggest rephrasing this sentence as it is both long and in which parts in brackets could be better integrated. Maybe something along the lines of:

To our knowledge, only one study attempts automating the assessment of building damage for volcanic hazards (Wang et al., 2024). In contrast, attention has been given to more commonly occurring hazards such as earthquakes and hurricanes, with the development of both mono- temporal (post-event imagery only) and multi-temporal (uses pre- and post-event imagery) approaches (Table 1).

In addition, some "multi-temporal" studies might also differ in the use of either a "before-after" approach (i.e., the comparison of two images) vs time-series approach. Maybe worth specifying if applicable to your problem.

- We have adopted both suggestions, the text now reads:

  To our knowledge, only one study attempts to automate the assessment of building damage for volcanic hazards (Wang et al., 2024). In contrast, attention has been given

to more commonly occurring hazards such as earthquakes and hurricanes, with the development of both mono- temporal (post-event imagery only) and multi-temporal (images taken at different times) approaches (Table 1). **Line 101-102:** I like the example! I might use a closer analogy to the problem tackled here, but I leave that to you.

- We have changed dogs and cats to:

  "between different roof types"

Line 135: Not wanting to open a can of worms - and totally aware of the use of post-disaster "opportunities", I find the use of the term "opportunity" perhaps a bit misplaced (i.e., using impacts on people's homes as the basis for research). I know this is not the case, but perhaps a more neutral phrasing would be more appropriate? (something along the lines of "prospect" - though I leave the selection of the most appropriate word to the native English-speaker authors).

- We agree that a more neutral phrasing would be better, we have changed this to:

  The 2021 eruption of La Soufrière volcano, St Vincent and the Grenadines, provided unprecedented circumstances for the collection of high-resolution UAV imagery

Line 205: Was the dataset manually geo-referenced? If yes - how?

- Images were not manually georeferenced, for model development we used the relative positions of bounding boxes within the images. This is now clarified as follows:

  Images do not contain GPS positioning or altitudes.

Line 213: This dataset does not contain any information regarding the flight path, which also raises the question regarding why buildings are captured with a lower resolution.

- This dataset (Dataset 2) does have GPS information as stated in original manuscript line 220. The drone was flown higher in this dataset than in the other two and buildings are visibly smaller, i.e. they cover less of the image frame.

Line 240: The reason behind this is not 100% clear

- Changed to:

  Given the absence of individual building location information, this number was approximated by overlaying Open Street Map building footprints with UAV GPS tracks where available.

Line 242: For off-nadir or very-off nadir images, how did you ensure that single bounding boxes did not overlap over buildings in the background?

- We have added in the following clarification:

  Boxes were drawn to fit the buildings closely and minimise background information. In areas where buildings are close together, off-nadir images may include parts of other buildings. Nevertheless, this was not considered an issue since deep learning models for object localisation will quickly learn to ignore background pixels

(Bouchard et al., 2022).

Line 254: That is not clearly intuitive given your description of the datasets. Did you filter out off- nadir images?

- We did not filter out off nadir images. We have removed and reworded this section, which now reads:

  The first tephra fall building damage state framework was developed after the eruption of Pinatubo, Philippines, 1991 (Spence et al., 1996), and was adapted from the macroseismic intensity scale used to evaluate seismic damage (Karnik et al., 1984). In the adapted framework damage ranges from DS0 – "no damage", through to DS5 – "complete roof collapse and severe damage to the rest of the building". Subsequent tephra fall building damage state frameworks were modified from the work of Spence et al., (1996) with changes in the wording made to reflect the characteristics of the case study (Table 2). In the damage state descriptions, damage to three critical aspects of a building is described: the roof covering, the roof structure, and the vertical structure (Blong 2003b; Hayes et al. 2019; Jenkins et al., 2024). In our study, most images depict buildings from an at nadir or close to nadir perspective making roof damage more discernible than damage to the vertical structure. Thus, we generated a damage state framework that is based on the proportion of observable damage to the roof, as in the work of Williams et al. (2020). Our final framework, which was developed over several iterations, classifies building damage into three classes: No observable damage to minor damage, Moderate damage, and Major damage (Table 3). Damage states are deliberately generic so that the range of possible damage to the range of different building types can be captured (Blong, 2003a). Our three classes are comparable to DS0-1, DS2, and DS3-5, respectively, of damage scales developed for ground surveys (Table 2).

Line 282/Section 2.3: This section is not the easiest to follow. For instance, from the first paragraph you mention splitting the classification task in two, a theme that you refer to in almost every paragraph, but without stating how or why. Can't this introduced and adequately presented in Section 2.3.3? Also, this section keeps on referring to sections ahead. Isn't it possible to optimise the writing to better integrate the development of the pipeline with its model components?

- We have revised the first paragraph to refer to only splitting the damage assessment into localisation and classification. The reference to splitting the classification into two is now moved down to the classification section. We now provide more rationale for our decision to split both aspects. The first paragraph of Section 2.3 now reads:

  After labelling, we split the full combined image dataset (2,811 frames from the UWI-TV, GOV and SRC sets) into train/validation/test sets (Figure 3). Given that a sizable proportion of the data did not contain GPS positions, images from each location were kept together to assure the train, validation, and test sets were independent. The partitioning was chosen to include diversity in both the image sets (UWI-TV/GOV/SRC) and in the location, which affects the thickness of tephra fall received. We aimed for a standard data split of 80/10/10, with the majority of data assigned for

training, however given the above constraints, this produced a split of 80% train, 8% validation, and 12% test (considering the number of bounding boxes and not the number of images). These datasets were used to develop our approach for building damage assessment. Most previous studies have split the damage assessment task into two subtasks: i) building localisation (i.e., identification of building bounding boxes within the images) and ii) damage classification (Table 1). While developing a model that can simultaneously locate and classify buildings with different levels of damage is feasible, model training under this approach can take significantly more time and resources to converge when compared to an approach that splits the tasks (Bouchard et al., 2022). Moreover, from an operational perspective, decoupling the two tasks makes the approach more flexible, for example, if building locations are already known then only the classification can be run, speeding up the remote assessment. For these reasons, we opted to split the building damage assessment task into two subtasks.

- The first paragraph of section 2.3.2 now reads:

  We chose to divide building damage classification into two separate classifications, Classifier 1 distinguishes between 'No damage to minor damage' versus the combined classes of 'Moderate damage' and 'Major damage', while Classifier 2 further differentiates between 'Moderate damage' and 'Major damage'. A hierarchical approach to classification has been found effective when the number of samples is limited or classes are unbalanced (Li et al., 2019b; An et al., 2021). We conducted experiments separately for Classifiers 1 and 2.

Line 292: Datasets?

- Changed

Line 294: In general, I personally recommend a more direct writing style, for instance changing: "we split the building damage assessment task into two subtasks, training and evaluating models for building localisation, which consists of identifying building bounding boxes within the images and building damage classification separately" to:

we split the building damage assessment task into two subtasks that include i) building localisation (i.e. identification of building bounding boxes within the images) and ii) building damage classification.

- We have edited this section as per the comment above (line 282). We have attempted to be more direct throughout, see revised Section 5.1 and excerpts from Section 3 above.

Line 299: This is true of most ML algorithms, not only deep learning

- Changed this to 'machine learning'.

Line 301: "experiment with different hyperparameter settings" or "optimise hyperparameters"?

- Changed as suggested.

Line 304: I don't think you need to state "(localisation, classification 1, classification 2)" at all in this

section, especially if just added in brackets. That makes reading heavy. (Same for line 309).

- Removed from both locations.

Line 309: Rephrase: "*Once we identified the best performing experimental setup for each task (building localisation, classification 1, classification 2), we combined the training and validation datasets and conducted K-fold cross-validation using the experimental setup and optimal hyperparameters that were identified (Cross validation: Section 3.1.3, Section 3.2.2)" To:*

After hyperparameter tuning, model accuracy was assessed using K-fold cross-validation (Cross validation: Section 3.1.3, Section 3.2.2)

- Changed to:

  Once we identified the best performing experimental setup for each task, we conducted K-fold cross validation to understand how the choice of training and validation data affects model performance (see Section 3.1.3, Section 3.2.2).

Line 316: In general, "data" is used in a very loose way. Would it work to change: "have data from more than one dataset"to: *contains images from more than one dataset*

- Changed.

Line 328: This statement is a bit out of place in a methodology section (plus it is somehow true for all contexts!)

- We have removed this sentence.

Line 331: Two comments here:

1. Following the comment on line 205, there seems to be georeferencing. This should be explained
2. Following the comment on line 242, the definition of the bounding boxes seems to be done on the orthomosaic? If yes, if I understand well, i) impact state is inferred from individual images ii) this is added as a label to the bounding boxes defined on the orthomosaic? This needs more clarity. I don't see any reference to the generation of the orthomosaic on Fig 3. The input to the model pipeline should probably be stated earlier.

   - We regret the confusion around the use of the orthomosaic image, bounding boxes were drawn on the non-georeferenced images and not on the orthomosaic. This has been clarified in response to general comment #2.

   - The generation of the orthomosaic is not included in the pipeline and should be done separately. We have now stated this both in Section 2.3 and in the caption for Figure 4. In Figure 4c, we mislabelled the orthomosaic 'Geotiff' which may have added to the confusion, we have now changed this. Figure 4 caption now reads:

     Figure 4. A schematic showing the full methodology for a) developing a model for building localisation, b) developing a sieve network, which acts as an add on to the building localisation model, c) developing a model for building damage classification and d) the building damage assessment pipeline developed in this work. The pipeline operates on an orthomosaic image (to be generated separately) and incorporates the final trained models for building localisation and two stages of building damage

classification along with all the necessary processing steps to link the models.

Line 349: Support opensource by citing the software used to produce the figure!

- Have added in the following to the figure caption:

  Pipeline generated using draw.io.

Line 362: Here again, unclear if "image feature" refers to individual images or the orthomosaic

- Here we are explaining generally how Faster R-CNN works, we have adapted the text to make this clear:

  Faster R-CNN is an improvement on the Fast R-CNN algorithm proposed by Girshick, (2015). The improvement comprises an initial region proposal network (RPN) which speeds up performance. In Faster R-CNN, image feature maps are extracted by passing the input image through a pretrained backbone CNN.

Line 377: Here - and anywhere else where you describe these "experiments", can you please specified how they were performed? Was it manually? In which case, is there any guidance on how you chose the ranges of each parameter? Or did you use optimisation algorithms?

- We have now included a more thorough description of what a single experiment consists of however we have left the ranges considered and their rationale for the supplementary material. Section 2.3.1 now reads:

  For object detection, to reduce model training and inference time, full sized images were split into image blocks. Experiments conducted as part of building localisation model selection included variations in the size of these blocks, the amount of overlap between blocks, and whether blocks were resized before training or not. We also experimented with the development of separate models for images captured with different viewing angles, training for only the SRC portion of the dataset (images mostly at nadir) and the combined UWI-TV-GOV portion (images mostly off-nadir). One experiment consisted of three replicates of a given combination of these aspects. Three replicates were conducted since the stochastic nature of the training process can cause models to converge at slightly different points (Aggarwal, 2018). For each experiment the replicate with the highest evaluation metric was the one compared against the other experiments. A total of 34 experiments were conducted to include all credible combinations of the varied hyperparameters and to find the best experimental setup for building localisation. More information on the hyperparameter values used in experiments can be found in the supplementary material.

  Section 2.3.2 now reads:

  For each experiment three replicates were conducted, each consisting of a grid search

to find the best combination of learning rate, batch size and L2 regularization. For more information on this see the supplementary material.

Line 381: This heading is inadequate (i.e., 2.3.1 is "Building localisation", why does 2.3.2. - and 2.3.3 too - need a "building"?). In addition, why not keeping these heading conceptual - e.g., "Building localisation" and "Building classification"? It seems to me that the sieve network is part of the building localisation task.

- We have removed the 'sieve network' heading throughout and text is now incorporated into the localisation section. We have changed building damage classification -> damage classification throughout.

Line 383: Define "small" or remove

- Removed.

Line 386: Rephrase: *The dataset used for training and evaluating the sieve network consists of randomly cropped background samples from full sized images in the training and validation sets*

- Changed as suggested.

Line 397: It seems that up to this point, the purposes of classifiers 1 and 2 have not been defined (unless we count Figure 3 as doing so). I might be mistaken, but I think this highlights the need of rethinking a bit the structure of Section 2.3.

- We have now explained in more detail the purpose of each classifier and the rationale for splitting the classification into two tasks:

  We chose to divide building damage classification into two separate classifications, Classifier 1 distinguishes between 'No damage to minor damage' versus the combined classes of 'Moderate damage' and 'Major damage', while Classifier 2 further differentiates between 'Moderate damage' and 'Major damage'. A hierarchical approach to classification has been found effective when the number of samples is limited or classes are unbalanced (Li et al., 2019b; An et al., 2021).

  Line 419: "False positive" has been used in line 385 but not defined

- We now define false positives in line 385:

  Bounding boxes produced by the detector are passed to the sieve network to filter out detections that are false positives. A false positive occurs when the detector predicts a bounding box that does not have an overlapping labelled building (i.e., detects a building when there is not one).

Line 444: *The five experiments with the highest average precision*

- Changed.

Line 455/Table 3: *Hyperparameters for the 5 experiments with highest average precision conducted for*

building...

- Changed.

By this point, the use of blocks vs boxes etc gets confusing for the reader. Maybe a conceptual sketch could help? Specifically, I don't think "block resizing" has been described in the text. I understand that a lot of the method is described in the SM, but the main m/s should be self- sufficient, therefore any concept shown in table/figures should be described in the text.

- In section 2.3.1 where we explain the building localisation method, we had previously mistakenly referred to the image blocks as patches. This has now been corrected, so this should make more sense now. Section 2.3.1 now reads:

  For object detection, to reduce model training and inference time, full sized images were split into image blocks. Experiments conducted as part of building localisation model selection included variations in the size of these blocks, the amount of overlap between blocks, and whether blocks were resized before training or not.

I suggest renaming the column "All training/ UWITV& GOV/ SRC" to "training dataset", assigning a letter to each dataset and defining it on a table footnote

- We have renamed the column as suggested.

Line 464: I think you are citing Table 6 before 5

- This has been rectified by adding the results described to the previous Table 4, and splitting what was previously Table 6 into two tables, one for localisation and one for classification and moving them to the associated sections. All tables are now cited in turn.

Line 519: Same as 444: *The five experiments with the highest macro F1 score*

- Changed.

Line 537: What do you mean by: "to understand the potential for our model to generalize to a new dataset"

- Changed to:

  … to understand how our model might perform on a new dataset.

Section 3: This section is very technical. Since the manuscript is rather oriented towards an impact/operational rather than a computer science audience, would it be possible to attempt better extracting what the raw values of model validation imply for a further application of the model? Discard this comment if you don't believe this is applicable.

- With this paper we were aiming to appeal to multiple audiences, striking a balance between the impact/operational perspective and the technical computer science side, with the majority of the technical results appearing in the supplementary material.

Nevertheless, we have reframed parts of the writing to be oriented towards an impact/operational as well as computing audience. We have added in further impact-oriented results and discussion through interpolation of the tephra isopachs and comparison with other existing works. We have added the following sentences to provide additional context for impact/ operational audiences. Hopefully the edits have made it clearer for all:

For the present application equal importance is given to: 1) making correct predictions about building locations, and 2) identifying as many buildings as possible. Consequently, striking the balance between precision and recall is crucial. We therefore selected the retrained detector + sieve network as the final building localisation model (Table 6).

To understand if a better model could be achieved with more data available for training, we combined the training and validation data and used this to retrain the best experimental setup for the detector. Evaluation of the retrained model on the test set resulted in an average precision increase from 0.701 to 0.751 for the non-sieved detector, and from 0.668 to 0.728 for the sieved detector, showing that having more data available for training produced a better model (Table 6).

Models trained to differentiate between the No damage to minor damage and Damaged classes performed better for the No damage to minor damage class, while those trained to differentiate between Moderate and Major damage performed better for the Major damage class (Table 7).

The tephra fall building damage map shown in Figure 7a was produced by overlaying the pipeline output georeferenced vector with the orthomosaic image in QGIS. Our remote damage assessment pipeline identified 442 buildings. Of these, 78% (N = 343) were classified as having No damage to minor damage, 9% (N = 40) as having Moderate damage and 13% (N = 59) as having Major damage. We observed that the two upper tephra fall thickness bins (70-80 mm and 80-90 mm), both had a higher proportion of buildings with Major damage compared to the lower thickness bins (Figure 7b, c), indicating a correlation between tephra fall thickness and building damage though it is not very pronounced. These findings are discussed in Section 5.3.

Line 598: I would rephrase to:

In order to optimise the application of separate models for building localisation and two stages of damage classification for operational contexts, we have integrated a damage assessment pipeline.

- We prefer to keep the initial phrasing.

Line 601: Again, here the use of orthomosaics is unclear. Also:

- Do you need to state these softwares? Isn't "computer vision" or "structure-from-motion" sufficient?
- Do you need to state "shapefile" - which is a proprietary file format? What about "georeferenced vector dataset"?

- Removed the list of potential softwares, and changed shapefile as suggested. The text now reads:

  The pipeline operates on an orthomosaic image and outputs a georeferenced vector set, with the following *attributes* for each building that is detected: *detection* (box confidence score), *ClassPred_1* (output class from Classifier 1, Damaged or No damage to minor damage), *ClassProb_1* (the probability of that class), *ClassPred_2* (output class from Classifier 2, Moderate damage or Major damage, this is only run if Classifier 1 outputs damage), *ClassProb_2* (the probability of the class output by Classifier 2), *damageState* (the final damage state).

Line 637-638: I don't understand this statement. What do you mean by distributions? Datasets? Building typology? If datasets, does it mean that your model is not generalisable?

Note: I see that some precisions are provided later. I still believe this should be clear from the beginning.

- Changed to:

  Deep learning models are known to perform well when the data they are evaluated on have similar characteristics to the data they were trained on, though have more difficulty when working with 'out of distribution' samples (Ben-David et al., 2010).