Response to Reviewer 2

Manuscript Title: *Climate change increases landslide susceptibility in Aotearoa New Zealand: Development and application of a national-scale model using machine learning*

Dear Editor and Anonymous Reviewer 2,

I would like to thank the anonymous reviewer for their thoughtful and constructive comments. In response, I have implemented several substantive improvements, including refinements to the model structure to improve physical realism and generalisation, and the addition of spatial cross-validation. I will also revise the manuscript to clarify key assumptions, limitations, and the interpretation of model outputs. These comments and the consequent changes have helped to strengthen the model its interpretation and presentation.

Note: Model and analysis updates described below have already been partially completed; textual revisions are described in terms of changes that will be made in the revised manuscript.

Sincerely,
Oliver Wigmore

Reviewer comments are shown in blue italics below, with author response in black.

General comments

The manuscript focuses on changes in national-scale rainfall-induced landslide susceptibility in New Zealand under the background of climate change. The topic has practical significance and is also within the scope of NHESS regarding natural hazards, climate change impacts, and risk assessment. Based on landslide data triggered by Cyclone Gabrielle, the author developed a LightGBM model and combined topographic, geological, soil, vegetation, and rainfall factors to predict landslide susceptibility under current and future SSP scenarios. Overall, the research idea is relatively clear, and the framework also has certain application value. However, the current manuscript still has several key methodological issues. The landslide inventory is mainly constructed based on NDVI-derived bare-ground change and a slope threshold, which introduces certain uncertainty in sample classification. The model training mainly relies on samples from the single Cyclone Gabrielle event and a limited region, but the model is then extended to the national scale and to different future storm scenarios, while its generalization ability has not been sufficiently validated. Important influencing factors such as antecedent soil moisture and aspect were not included. The resampling of HIRDS rainfall data from 2000 m to 25 m may also introduce scale mismatch and spatial bias. In addition, although the ROC-AUC is high, the PR-AUC for the landslide class is relatively low, and at the 5% threshold, false positives are clearly more numerous than true positives, indicating that the model has a clear tendency toward over-prediction. Therefore, I believe that the current results are not sufficient to support the main conclusions of the manuscript regarding changes in landslide susceptibility at the national scale and under future climate scenarios. The author is advised to consider resubmission only after further improving the landslide inventory, adding validation using independent events or independent regions, conducting sensitivity analyses of thresholds and *prior probabilities, and fully discussing the uncertainty of future rainfall data. For the above reasons, I recommend that the current manuscript not be accepted.*

Response:

I thank the reviewer for their thoughtful comments regarding model generalisation, training data limitations, transferability, and interpretation of the national-scale results.

I note that suitable independent validation outside the Cyclone Gabrielle training region is not readily available and agree that this represents an important limitation of the study. Consequently, the absolute probabilities of landslide occurrence predicted by the model cannot be assumed to be perfectly calibrated throughout New Zealand, particularly in regions where geomorphic processes may differ from those represented in the training data.

I acknowledge that the absence of independent validation limits confidence in model transferability; however, it does not necessarily imply an absence of model utility or discrimination outside the training domain. The model is not intended to provide site-specific forecasts of landslide occurrence, but rather to identify relative spatial patterns of susceptibility under present and future rainfall scenarios. While uncertainty in probability calibration is expected to increase outside the training region, the ability of the model to distinguish relatively more susceptible terrain from relatively less susceptible terrain is likely to be more robust. Because it is based on predictors representing fundamental controls on shallow rainfall-induced landslides, including slope, rainfall, vegetation cover, soil properties, and critical-zone architecture (e.g. rooting depth, particle size, depth to zone of slow permeability, depth to hydrogeologic basement, and rock density). Furthermore, these predictor variables were specifically derived and selected to improve transferability.

To strengthen the treatment of model generalisation, I have substantially revised the modelling framework. The revised model incorporates:

- spatially blocked cross-validation (5 km blocks).
- a more constrained hyperparameter search space to improve regularisation and reduce overfitting.
- monotonic constraints on rainfall predictors to ensure physically consistent model behaviour under increasing rainfall conditions.

Under this revised framework:

- holdout ROC-AUC is 0.92.
- spatially blocked cross-validation ROC-AUC is 0.90.

These results suggest that model performance remains robust when evaluated under validation approaches specifically designed to account for spatial autocorrelation and reduce the risk of optimistic performance estimates.

I also note that Appendix C demonstrates that predictor values across New Zealand fall within the range represented by the training dataset. Therefore, the model is not extrapolating beyond the numerical range of predictor values used during training. Nevertheless, I acknowledge that this does not guarantee transferability of the predictor–response relationships themselves. Regions may differ in the physical processes governing landslide occurrence, such that relationships learned in the Cyclone Gabrielle region may not transfer perfectly elsewhere.

This is likely to have a greater influence on the calibration of absolute RIL probabilities than on the relative ranking of terrain susceptibility. Many of the fundamental controls on shallow rainfall-induced landsliding (e.g. slope, rainfall magnitude, vegetation cover, and critical-zone properties) are common across regions, even though the precise thresholds and process responses associated with failure are likely to vary spatially. Consequently, absolute susceptibility estimates should be interpreted with greater caution than relative susceptibility patterns.

To further evaluate the robustness of the national-scale conclusions, I conducted sensitivity analyses using alternative methods of translating susceptibility outputs into impacted-area estimates, including threshold-independent probability aggregation and confusion-matrix error propagation approaches. While the absolute magnitude of impacts varied between methods, the direction of change and identification of the principal hotspot regions remained broadly consistent. This suggests that uncertainty associated with model transferability primarily affects the magnitude and calibration of predictions, rather than the overall spatial pattern of elevated susceptibility.

In response to the reviewer's concerns, I will revise the manuscript to more clearly position the model as a scenario-based, screening-level susceptibility framework informed by a well-characterised extreme event, rather than a fully validated predictive model applicable to all geomorphic settings. I will also explicitly state that:

- the model is trained within the Cyclone Gabrielle region and subsequently applied nationally.
- confidence in absolute probabilities decreases outside the training domain.
- results in regions with substantially different geomorphic and hydroclimatic conditions (particularly the Southern Alps and Fiordland) should be interpreted with caution.
- the primary value of the modelling framework lies in identifying relative changes in susceptibility under different rainfall scenarios, rather than providing precise forecasts of future landslide occurrence.

I believe these revisions substantially strengthen the treatment of model generalisation, uncertainty, and transferability, while more clearly communicating the intended interpretation and limitations of the national-scale results.

Specific comments

1. *The rationale for the construction of the landslide inventory and the selection of thresholds needs to be further explained. In the manuscript, Sentinel-2 images before and after Cyclone Gabrielle are used to identify bare-ground changes through NDVI changes, and then a slope >10° rule is applied to distinguish landslides from flood deposits and other areas. This approach is reasonable to some extent, but NDVI essentially identifies bare-ground change rather than landslides themselves. At the same time, the slope >10° threshold is an empirical threshold, which may affect which areas are ultimately defined as landslide samples. The author is advised to explain the basis for selecting the NDVI threshold and the slope threshold, and to further discuss whether different slope thresholds would lead to obvious changes in the number of landslide samples, their spatial distribution, and the model results.*

Response:

I acknowledge the reviewer's concerns regarding the use of NDVI-based bare-ground detection and empirical thresholds.

I will:

- expand the justification for the ΔNDVI threshold selection;
- clarify the rationale for the slope >10° filter (to separate hillslope failures from floodplain deposition).
- include additional discussion of potential sensitivity to threshold choice.

I will also clarify that:

- the short temporal window (~2 weeks) reduces the likelihood of non-landslide (bare-ground transition) land-cover change.
- combining two independent inventories increases label reliability, although it may reduce completeness.
- this represents a standard and widely adopted approach in remote sensing-based landslide mapping, where trade-offs between completeness and classification reliability are unavoidable.

1. *The author combined the landslide data identified in this study with the landslide inventory from Dragonfly Data Science, and only used pixels identified as landslides in both datasets as positive samples. Areas identified by only one of the two datasets were classified as "possible landslides" and excluded from both training and testing. This approach can reduce misclassification, but it may also make the training samples more biased toward relatively obvious and larger landslides, causing the model to be more capable of identifying large areas of bare-ground change rather than necessarily representing all rainfall-induced landslides. The author is advised to further explain whether this treatment affects the completeness of the landslide samples, and whether excluding the "possible landslide" areas changes the spatial representativeness of the non-landslide samples.*

Response:

I thank the reviewer for raising this point.

I agree that prioritising agreement between inventories may reduce completeness and favour higher-confidence mapped failures, in that only areas consistently identified across both inventories are retained. However, because both inventories were independently derived from Sentinel-2 imagery at the same 10 m spatial resolution and utilise pixel-level classifications, I do not expect this effect to be strongly related to landslide size. The inventory construction process does not impose any additional minimum landslide size threshold beyond that inherent in the source imagery itself (which is then aggregated to the 25m model grid pixels). Rather, the principal effect of intersecting the inventories is to prioritise label reliability over inventory completeness.

This was a deliberate methodological choice. In machine learning models, particularly under strong class imbalance, higher-confidence labels are generally preferred over larger but noisier datasets, as label noise can degrade model performance and obscure underlying relationships. Importantly, the resulting training dataset remains large (~370,000 pixels, 50:50 balanced class prevalence), ensuring robust model learning. This approach therefore strengthens the quality of the predictor–response relationship used for model training, which is critical for robust machine learning performance.

I will revise the manuscript to:

- explicitly state that the dataset prioritises label reliability over completeness.
- clarify that this filtering applies only to training/testing data construction, and does not limit the spatial applicability of the model outputs.

2. *Regarding the representativeness of the training area and the issue of national-scale extrapolation, the author is advised to provide further justification. The model is mainly trained using landslide samples from the area affected by Cyclone Gabrielle, but it is ultimately applied to the whole of New Zealand. Considering that different regions may vary in terms of topography, geology, land cover, and rainfall characteristics, whether a model trained only on this event and this region is sufficient to support national-scale prediction needs to be more fully explained.*

Response:

I acknowledge that independent validation across different geographic regions is not available and represents a limitation of this study.

However, I note that many data-driven landslide susceptibility studies rely on a single well-characterised event or region, particularly where detailed landslide inventories and event-specific trigger data are required.

To strengthen the treatment of generalisation, I will:

- incorporate spatially blocked cross-validation, providing an explicit assessment of generalisation within the training domain.
- clarify that the model is not extrapolating beyond the range of predictor values seen during training, as demonstrated by comparison of predictor distributions across New Zealand (Appendix C).
- explicitly state that predictions outside the training domain should be interpreted as screening-level estimates, because the underlying physical processes that drive RIL susceptibility may differ outside the training region.
- Include a discussion of where and why model results are likely to be over/under predictive – e.g. highlighting the West Coast and Fiordland (extreme rainfall areas).

In addition, I will emphasise that the model estimates: relative spatial patterns of susceptibility under defined rainfall scenarios, rather than universally generalisable predictions across all geomorphic settings.

This level of interpretation is consistent with many national-scale landslide susceptibility studies, where models trained on well-characterised events are applied more broadly to identify relative patterns of risk.

I also note that while independent validation outside the training region is not available, the absence of such validation does not in itself demonstrate a loss of discriminative ability. Rather, it limits confidence in the calibration of absolute probabilities and susceptibility magnitudes. Accordingly, the model outputs are interpreted primarily as relative indicators of susceptibility under alternative rainfall scenarios, rather than precisely calibrated regional forecasts of landslide occurrence.

3. *The model training samples come from the single event of Cyclone Gabrielle. This event has its own rainfall path, duration, storm direction, antecedent wetness conditions, and regional spatial characteristics. Therefore, whether the model can be generalized to other rainfall events still needs further discussion. If there is no validation using other independent rainfall events, the author is advised to be more cautious when describing the applicability and generalization ability of the model.*

Response:

I acknowledge that the model is trained on a single event, and that this is a potential limit of the study.

I will clarify that:

- the model is event-conditioned, capturing relationships observed during Cyclone Gabrielle.
- it does not represent a fully generalised multi-event model.

I also note that:

- the training dataset spans a wide range of rainfall intensities and environmental conditions, effectively providing a broad training domain within a single event.
- the revised modelling framework (including spatial cross-validation and increased regularisation) demonstrates that the model captures stable relationships rather than highly localised artefacts.

4. *Rainfall-induced landslides are usually affected not only by event rainfall, but also closely related to pre-storm soil moisture conditions. The author is advised to further discuss the possible impacts of the absence of antecedent soil moisture, and to consider whether antecedent rainfall indices, soil moisture reanalysis data, or other data could be used for supplementary testing or sensitivity analysis.*

I agree that antecedent soil moisture is an important control on rainfall-induced landslides.

I will expand the manuscript to clarify that antecedent soil moisture is not included because it cannot be robustly specified for future storm scenarios, which are a primary focus of this study. Including such a variable would substantially reduce the transferability of the framework to projected future rainfall scenarios.

I will also clarify that the training event (Cyclone Gabrielle) occurred following sustained antecedent rainfall, so soil moisture conditions in the training region were already elevated. This means the model is trained under a relatively high preconditioning state, which may approximate the upper range of susceptibility for a given rainfall magnitude.

At the same time, recent analysis of Cyclone Gabrielle landslides by Massey et al. (2025) suggests that, in some areas, 24-h rainfall totals were sufficiently high to “overprint” antecedent soil water conditions, such that pre-event soil moisture may not have materially influenced the spatial landslide distribution. I will therefore clarify that the omission of antecedent soil moisture is less problematic for this particular extreme training event than it would be for more moderate storms, but still remains a limitation when interpreting the model more generally.

Accordingly, the revised manuscript will state that:

- the model likely represents susceptibility under a relatively high antecedent moisture state;
- susceptibility may be overestimated under drier antecedent conditions; but
- for very high-intensity rainfall events such as Cyclone Gabrielle, antecedent soil moisture may have been partly overprinted by the magnitude of the event rainfall itself.

5. *The author did not include the aspect variable, on the grounds that the influence of aspect on landslides may depend on storm direction, while future storm direction is unknown. This explanation is understandable, but for the specific training event of Cyclone Gabrielle, aspect and storm direction may still have influenced the landslide distribution. The fact that storm direction cannot be determined in future scenarios does not mean that aspect is completely unimportant in the current event-based modelling. The author is advised to further explain the possible influence of excluding the aspect variable on model interpretation and prediction results.*

Response:

I thank the reviewer for this comment and agree that aspect may have influenced landslide occurrence during Cyclone Gabrielle due to storm direction.

However, the decision to exclude aspect as a predictor was made deliberately to improve model transferability beyond the training event. Aspect effects are strongly dependent on storm direction, and including aspect would likely encode event-specific directional biases, improving model fit to the training event but reducing its ability to generalise to other storm scenarios, particularly future conditions where storm direction is unknown.

I note that excluding aspect does not prevent the model from capturing related behaviour indirectly through other predictors, but avoids explicitly embedding direction-dependent relationships in the model structure.

In this context, excluding aspect helps avoid embedding non-transferable directional dependencies in the model structure.

I will revise the manuscript to more clearly explain this trade-off between event-specific performance and general applicability.

6. *Regarding the accuracy and applicability of the future precipitation data itself, the author is advised to provide further explanation. Precipitation has significant spatial heterogeneity,*

and whether HIRDS data are reliable in representing the spatial differences in future precipitation needs to be clarified. In addition, the original resolution is 2000 m, and it is resampled to 25 m; this process may further introduce spatial bias. The author is advised to discuss whether different resampling methods, such as nearest neighbour, cubic convolution, and inverse distance weighting, would have a significant impact on the landslide susceptibility prediction results.

Response:

I thank the reviewer for raising this point.

I note that a recent study by Sigid et al., (2026) investigated changes in and spatial heterogeneity of future rainfall under climate change. They found that *"spatial heterogeneity in projected changes to both the intensity and frequency of extreme rainfall suggests that identifying regionally differentiated changes in extreme rainfall may be inherently challenging".* They suggested that because of this, approaches that apply uniform nationwide percentage changes as in the HIRDs dataset remain a reasonable approach. However, I will ensure that the limitations of this approach are clearly communicated and include reference to Sigid et al., in the manuscript to support the use of HIRDS data in this study.

I agree that resampling rainfall from the HIRDS grid (~2000 m) to the 25 m modelling grid introduces uncertainty and does not capture fine-scale spatial variability, as discussed in the manuscript.

While different interpolation methods (e.g. nearest neighbour, bilinear, cubic convolution, inverse distance weighting) may produce slightly different spatial patterns, I note that at this scale transition the dominant source of uncertainty is not the choice of interpolation method, but the assumption that rainfall fields can be statistically downscaled without explicit representation of physical processes, particularly topographic controls on precipitation.

Capturing such processes would require physically based or dynamically downscaled rainfall products, which are not currently available at a national scale and are beyond the scope of this study.

As such, differences between interpolation methods are expected to be secondary relative to this broader limitation.

I will revise the manuscript to:

- reference the Sigid et al, (2026) study.
- clarify the limitations of statistical downscaling.
- distinguish between uncertainties arising from interpolation choice and those arising from scale mismatch and lack of physical downscaling.
- emphasise that rainfall inputs represent broad-scale spatial patterns, consistent with national-scale susceptibility modelling.

7. *The model results show an increase in landslide susceptibility under future warming scenarios, and this direction is reasonable. However, the magnitude of the increase may be affected by factors such as the probability threshold and the prior probability setting. The*

author is advised to further explain whether the magnitude of the future increase in landslide susceptibility remains stable under different thresholds or different prior probabilities.

Response

I thank the reviewer for this important comment. I agree that the magnitude of threshold-based outputs, including mapped susceptible area and exposure estimates, is sensitive to the selected probability threshold.

I would like to clarify, however, that the prior-probability adjustment applied in this study functions as a monotonic scaling of the raw model probabilities. As a result, it does not alter the ranking of pixels, and therefore does not affect discrimination metrics such as ROC-AUC or PR-AUC. Instead, it changes the absolute probability scale, which in turn affects how many pixels exceed a given fixed threshold and are therefore mapped as susceptible or counted in exposure analyses.

I also agree that threshold choice has an important influence on the absolute magnitude of mapped susceptible area and exposure estimates. In the revised manuscript, I will therefore make clearer that thresholded outputs should be interpreted as comparative indicators across scenarios, rather than precise quantitative forecasts.

At the same time, I note that the affected-area plots based on confusion-matrix error propagation are less sensitive to threshold choice than the raw thresholded area maps, because they partially correct for the threshold-dependent balance between false positives and false negatives (assuming that the error rates remain stable across the prediction domain). This correction does not remove threshold dependence entirely, but it does reduce its influence and provides a more stable basis for comparing scenarios.

Similarly comparing percentage changes in susceptible area (or TP-FN corrected area), is less affected by threshold selection.

To further investigate the robustness of the national-scale conclusions I have undertaken a sensitivity analysis of the threshold analysis. This included comparing regional metrics using a 1% and 5% probability threshold as well as the calculation of a threshold-independent probability aggregation approach. While the absolute magnitude of impacts varied between methods, the direction of change and identification of the principal hotspot regions remained broadly consistent. This suggests that uncertainty associated with model transferability primarily affects the magnitude and calibration of predictions, rather than the overall spatial pattern of elevated susceptibility.

Accordingly, I will revise the manuscript to explicitly state that:

- the ranking of susceptibility is unaffected by the prior-probability adjustment.
- the absolute area and exposure values derived from fixed thresholds are sensitive to threshold selection and probability scaling, but that the general findings remain consistent.
- the error-propagated area estimates reduce, but do not eliminate, this threshold sensitivity.
- Include the core findings of the sensitivity analysis.

- adjust figures so that they focus more on percentage changes in susceptible area – as opposed to absolute area values.
- conclusions are therefore based primarily on the relative ordering of scenarios and the consistency of broad spatial patterns, rather than on any single absolute estimate of impacted area or exposed assets.

I will also expand the Results and Discussion to make this distinction clearer for the reader.

8. *Although the ROC-AUC is high, the PR-AUC for the landslide class is relatively low, and at the 5% threshold, false positives far outnumber true positives. The author is advised to make this point clearer in the interpretation of the results, and to further discuss the possible effects of this over-prediction on the landslide susceptibility maps, the estimated affected area, and the exposure assessment of buildings and roads.*

Response:
I thank the reviewer for highlighting this important point. I agree that the high absolute number of false positives at fixed probability thresholds requires careful interpretation, particularly for the susceptibility maps and the downstream analyses of affected area, buildings, and roads.

The manuscript reports a ROC-AUC of 0.94 and a PR-AUC of 0.276 for the landslide class. These metrics evaluate different aspects of model performance. ROC-AUC measures the ability of the model to rank landslide pixels above non-landslide pixels and is relatively insensitive to class prevalence, whereas PR-AUC evaluates the precision–recall trade-off for the positive (landslide) class and is strongly influenced by the prevalence of landslides within the evaluation dataset (Richardson et al., 2024).

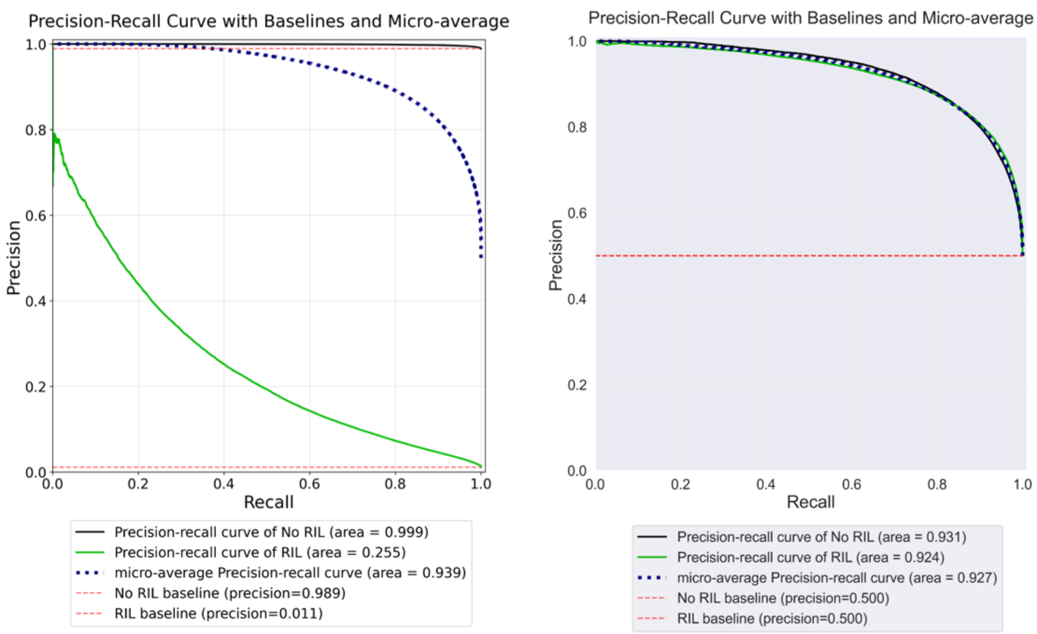
In this study, the holdout test dataset was constructed to reflect the observed event prevalence (~99:1 no-RIL:RIL). Under these conditions, the baseline precision is approximately 0.01, meaning that a random classifier would be expected to achieve a PR-AUC close to 0.01. Consequently, a PR-AUC of 0.276 represents approximately a 26-fold improvement over the baseline classifier and indicates substantially stronger discriminative performance than the absolute value alone may suggest.

Because many landslide susceptibility studies evaluate PR-AUC on balanced or artificially rebalanced datasets, I have also calculated the PR-AUC on a 50:50 subset of the same holdout test dataset. This produced a PR-AUC of 0.92. I do not consider this to be a realistic estimate of deployment performance, nor do I propose it as an alternative measure of model accuracy. Rather, it is presented solely to illustrate the strong dependence of PR-AUC on class prevalence. The underlying classifier is unchanged; only the prevalence of the evaluation dataset differs. I therefore include both the balanced and true-prevalence PR curves in the Appendix for interpretive context.

I nevertheless agree with the reviewer's broader point that applying fixed probability thresholds to a rare-event problem results in substantial numbers of false positives. Even relatively low false-positive rates can produce large numbers of falsely classified pixels when the negative class is very large. Consequently, thresholded susceptibility maps should be interpreted as conservative screening products rather than deterministic forecasts of landslide occurrence.

In the revised manuscript I will therefore:

- clarify that PR-AUC should be interpreted relative to class prevalence.
- include the balanced PR curves in the Appendix for interpretive context, and retain the 99:1 (true-prevalence) PR curve in the main text.
- expand discussion of the implications of false positives for susceptibility mapping and exposure analysis.
- more clearly state that thresholded susceptibility maps are intended as screening-level products rather than precise forecasts of landslide occurrence.



PR-AUC curves on a 99:1 test set (left) and a 50:50 balanced test set (right). Note these are for the updated model.

References:

Massey, C., Leith, K., Robinson, T. R., Lukovic, B., McColl, S., Carey-Smith, T., Rosser, B., Wotherspoon, L., Smith, H., Betts, H., Buxton, R., & Bidmead, J. (2025). What controlled the occurrence of more than 116,000 human-mapped landslides triggered by Cyclone Gabrielle, New Zealand? *Landslides*. https://doi.org/10.1007/s10346-025-02591-y

Richardson, Eve, Raphael Trevizani, Jason A. Greenbaum, Hannah Carter, Morten Nielsen, and Bjoern Peters. “The Receiver Operating Characteristic Curve Accurately Assesses Imbalanced Datasets.” *Patterns* 5, no. 6 (2024): 100994. https://doi.org/10.1016/j.patter.2024.100994.

Sigid, M. F., Harrington, L. J., & Lewis, H. (2026). On the Future of Extreme Rainfall in New Zealand. *Earth’s Future*, *14*(4), e2025EF007427. https://doi.org/10.1029/2025EF007427