Response to Reviewer 1

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 1,

I would like to thank the 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.

The manuscript focuses on the development of a national-scale rainfall-induced landslide model for New Zealand and the projected changes in landslide susceptibility under future climate scenarios. The topic is of clear practical relevance. The authors attempt to incorporate triggering rainfall into the model and further combine the modelling results with future climate scenarios and exposure analyses for roads and buildings, thereby building a national-scale application framework. From a methodological perspective, the manuscript include landslide sample extraction, predictor selection, LightGBM modelling, SHAP interpretation, and scenario-based projection. The amount of work is substantial, and the results have reference value. However, there are several important issues still require further clarification.

1) *The training data are based on landslides triggered by Cyclone Gabrielle in Hawke's Bay and Gisborne/Tairāwhiti. The authors also acknowledge that the South Island, especially the Southern Alps, differs substantially from the training area in both lithology and dominant slope processes. Pages 21-22 further state that the South Island is dominated by harder metamorphic rocks, and that the Southern Alps commonly feature scree/talus slopes, rock avalanches, and large deep-seated landslides, which are not equal to the shallow rainfall-induced landslides that are more typical of the North Island. The manuscript also notes that "the interpretation of RIL susceptibility maps for the Southern Alps should be undertaken cautiously." It is suggested that the authors either state the applicable scope more accurately in the title and main conclusions, or add cross-regional validation.*

RESPONSE:

I agree that applying a model trained on a single event and region to the national scale introduces limitations, particularly given the contrasting geomorphic processes and lithologies of the Southern Alps.

While cross-regional validation would be desirable, suitable independent landslide inventories for regions such as the Southern Alps are not currently available.

To improve transferability I was careful to not rely on geologic class directly, which cannot be implemented in areas that it was not trained on. Instead, I have selected variables that aim to describe the architecture of the critical zone (e.g. rooting depth, particle size, depth to zone of slow permeability, depth to hydrogeologic basement, and rock density) i.e. the sediments within which most RIL occur. I note that Appendix 2 demonstrates that predictor values in application areas do not exceed the range represented in the training data, so the model is not extrapolating beyond its training range.

However, I acknowledge (and agree with the reviewer) that this does not account for the likely reality that the underlying physical processes that drive RIL occurrence in different regions may be different, i.e. the relationships between the predictor variables that are learned in the training region may not transfer to other regions.

To address this concern, I will revise the manuscript to explicitly clarify the model's scope. In particular, I will:

- Modify the title and conclusions to clearly state that the model is regionally trained but applied at national scale.
- Emphasise that interpretation outside the training domain, especially in the Southern Alps, should be undertaken with caution (and what that means) due to differing landslide processes.
- Provide clearer discussion of where the model is likely to under/over represent susceptibility based on this.

Additionally, I have implemented structural changes to the model that improve its generalisation ability. These are described in my response to reviewer point 4) below.

2) *The authors identify bare-ground change using a ΔNDVI threshold derived from pre- and post-event multi-temporal Sentinel-2 imagery, and then filter landslide pixels using a slope threshold of >10°. As explicitly stated on page 5, this approach "includes both the initiation point and run-out zone." This means that the model is not learning purely the conditions of landslide initiation, but rather the spatial pattern of terrain affected after the event. On page 20, the authors also state that the model "does not separate the RIL initiation point from the debris run-out zone." This has direct implications for the physical interpretation of predictors such as slope, curvature, TWI, and soil depth, because the geomorphic characteristics of runout zones differ from those of initiation zones. The authors should clarify more explicitly whether the model output is closer to initiation susceptibility, event-affected area probability, or a mixed footprint probability.*

RESPONSE:

I agree that deriving landslide inventories from bare-ground change results in the inclusion of both initiation zones and runout areas.

Because of this the model does not learn purely initiation conditions. Instead, it captures the spatial footprint of landsliding as expressed in the post-event landscape. While the model does not explicitly simulate runout processes, some predictors (e.g. curvature and TWI) implicitly incorporate neighbourhood-scale information that enables partial representation of runout zones. Consequently, the model output should not be interpreted as representing purely landslide initiation susceptibility.

I will revise the manuscript to explicitly clarify that:

- The model output represents a mixed footprint probability, incorporating both initiation zones and, in many cases, portions of runout areas
- This has implications for the interpretation of terrain-related predictors (e.g. slope, curvature, TWI), which may reflect both initiation and deposition environments

3) *The authors use slope >10° to distinguish landslide-related bare-ground change from non-landslide areas. On page 13, in the SHAP interpretation, the manuscript itself states that the lower susceptibility below 10° "may also be a consequence of the RIL mapping exercise." This indicates that part of the slope effect is directly embedded in the sample construction procedure, rather than being entirely learned by the model from independent observations of the process. Since slope is identified as the second most important predictor, this issue should be explained more carefully in the Results and Discussion.*

RESPONSE:

I agree that the use of a slope threshold (>10°) in constructing the landslide inventory introduces a degree of structural bias into the model.

This threshold was necessary to distinguish landslide-related bare-ground change from extensive flood-induced sediment deposition associated with Cyclone Gabrielle, which produced widespread bare-ground signals in low-gradient valley floors. Consequently, part of the importance of slope in the model reflects this preprocessing decision and should not be interpreted as a purely independent learned control.

We will revise the manuscript to:

- Provide a clearer justification for the slope threshold applied during inventory construction
- Explicitly acknowledge that this threshold partially influences the learned relationship between slope and susceptibility
- Discuss how this should be considered when interpreting slope importance in the SHAP analysis

4) *The authors first hold out 20% of the landslide polygons, then randomly sample non-landslide pixels from the full region, and use a 10 m exclusion buffer and a 72 m buffer to reduce spatial leakage. Even so, the training and testing data still come from the same Cyclone Gabrielle event and the same broader regional context. In addition, the 10-fold cross-validation used in the manuscript is random cross-validation and does not explicitly account for spatial autocorrelation. It is suggested that the authors further evaluate the model using spatial cross-validation.*

RESPONSE:

I have implemented substantial updates to address concerns regarding spatial dependence and generalisation.

Model architecture improvements

To improve generalisation and reduce possible overfitting:

- The Optuna search space has been more tightly constrained, resulting in shallower trees and reduced model complexity, with only a small drop in ROC-AUC (from 0.94 to 0.92) on holdout test data. Changes in the other model metrics are similarly modest.
- Positive monotonic constraints have been applied to rainfall predictors (total rainfall and rainfall as a proportion of annual rainfall), ensuring physically consistent behaviour (i.e. increasing rainfall always increases landslide susceptibility)

These changes improve both interpretability and robustness, and reduce the likelihood that the model learns event-specific rainfall patterns.

However, it is important to note that the updated model with enforced monotonicity prescribes the direction of the rainfall–susceptibility relationship (i.e. increasing rainfall results in increased susceptibility). As such, this relationship should now be interpreted as a physically constrained modelling assumption rather than an independently learned causal relationship. This clarification will be made explicit in the revised manuscript.

Spatial cross-validation

We have also implemented two forms of spatial cross-validation using a 5 km grid:

1. Spatially blocked 10-fold cross-validation on the training dataset (50:50 prevalence) to assess model stability
2. Spatially independent evaluation of predictive performance on holdout data (99:1 prevalence) using equivalent (held-out) spatial blocks

This analysis shows only a modest reduction in ROC-AUC (from 0.92 to 0.90), indicating that model performance is not substantially inflated by spatial autocorrelation.

Summary

Together, these changes:

- Improve model generalisation
- Increase physical realism

- Provide stronger validation of predictive performance in spatially independent contexts

These updates will be fully described in the revised manuscript.

5) *The future scenarios are derived from HIRDS rainfall data at 2000 m resolution, statistically downscaled to 25 m, and then combined with the change factors in Table 2 and the national mean temperature changes in Table 3 to construct the SSP2-4.5 and SSP3-7.0 scenarios. On the one hand, the future intensification signal is based on national mean warming rather than finer-scale regional climate differences. On the other hand, the manuscript does not systematically report the uncertainty range across the six climate models, but instead directly uses the national mean warming signal. The resulting outputs can be considered as national-scale scenario projections, but they may not be sufficient to support more detailed local-scale interpretations. It is suggested that the authors add corresponding statements or uncertainty discussion in the Abstract, Discussion, and Conclusion.*

RESPONSE:

I agree that the use of nationally averaged climate signals and the absence of explicitly quantified inter-model uncertainty could limit the interpretation of results at finer spatial scales.

We will revise the manuscript to:

- Clearly state that model outputs should be interpreted as national-scale scenario projections
- Emphasise that they are not intended for detailed local-scale application
- Expand discussion of uncertainty in the Abstract, Discussion, and Conclusions, including both climate model variability and downscaling limitations

I note that applying nationally uniform change factors is consistent with established approaches in New Zealand (e.g. HIRDS). Recent work by Sigid et al., (2026) highlights the challenges in robustly quantifying spatial variability in extreme rainfall projections, and consequently that the application of national scale percentage change factors likely remains reasonable. However, we will ensure that the limitations of this approach are clearly communicated.

Sigid, M. F., Harrington, L. J., & Lewis, H. (2026). On the future of extreme rainfall in New Zealand. Earth's Future, 14, e2025EF007427. https://doi.org/10.1029/2025EF007427

6) *The presentation of the SSP scenarios should be made consistent throughout the manuscript. It would be better to use SSP2-4.5 and SSP3-7.0 consistently, rather than 2-4.5 and 3-7.0.*

RESPONSE:

Agreed, I will revise the manuscript to consistently use SSP2-4.5 and SSP3-7.0 throughout.