Complementary responses (R) to the RC2 comments:

1. What precisely is the prediction target?
The stated objective is to predict slope instability, but the output variable is a weighted aggregation of confirmed and unconfirmed emergency calls. These calls may reflect physical instability, population density, reporting behaviour, access to municipal services, and public awareness. Please clarify whether the model is intended to forecast: (i) verified mass-movement occurrence, (ii) emergency-call intensity, or (iii) a broader municipal-response indicator. The terminology, evaluation metrics, and practical claims should be aligned with the actual target.

R – The model is intended to forecast emergency-call intensity/density, which, in turn, is closely related to verified mass-movement occurrences at different scales. Please note that only emergency calls related to the subject mass movement are selected. As explained in the previous answer for RC1, the output variable (emergency calls to the municipality) reflects human reporting behavior in response to risk exposure and does not always reflect actual physical instability. This is considered in the paper by the RFU, the ratio failure/unconfirmed call (RFU) weight: "In order to take this into consideration in the analysis, confirmed calls were given a higher weight (RFU value) in the output variable value, which is calculated as the weighted sum of all the calls in the adopted time interval." All citizens are supposed to know the procedure for calling the municipality, and no formal procedure is required; just a phone app.

2. How can the model be applied citywide if slopes without emergency calls are excluded? The manuscript defines an eligible slope as a slope with at least one emergency call within its area of influence. This preselection removes true negative cases. A model trained only on call-positive locations cannot demonstrate its ability to distinguish unstable slopes from stable slopes across the city. Please reconstruct the modelling dataset to include slopes with no calls and time periods without recorded events, or limit the stated objective to ranking call intensity within previously identified risk areas.

R – The slopes areas with no calls contain only stable slopes, and the research focus is on risk areas with the occurrence of mass movements triggered in the raining periods. Furthermore, the number of slopes with no calls associated in each rainy day is much higher than those inside the vicinity of a call, and is difficult to know how far we are from failure with no calls. The model was designed to forecast and declair warnings and emergencies, preventing economic and human life losses. Therefore, in the authors' opinion, there is no limitation to the model's interpretation because stable slopes and non-risk areas are not the main concern of the practical problem.

3. How have spatial and temporal information leakage been prevented?
Approximately two million derived observations are generated from overlapping slope areas, emergency calls, radii of influence, and time windows. Randomly distributing these observations among training, validation, and testing subsets is likely to place highly related records from the same slope, neighbourhood, rainfall episode, or emergency-call cluster into different subsets.
This may substantially inflate the reported R^2 values. Please adopt grouped spatial and temporal validation, such as spatial-block cross-validation and an out-of-time holdout period. Hyperparameter and scenario selection must use only the training and validation subsets; the testing subset should be evaluated once after all decisions are finalized.

R - The number of observations changed with the radius of influence. It means that, for each call, the number of slopes falling within the influence area changed with the radius used in the analysis: 75, 100, 150, 200, and 300m. The following text was added to the paper: “Increasing RFU, Ti, and Ri also increases the value of the output variable. Analysis with a higher Ri will encompass a larger

area and involve more slopes per emergency call. The combinations of slopes and emergence calls for the period of analysis yielded approximately 872,000 and 11,000,000 observations for Ri values of 75m and 300m, respectively, for each RFU, Ti, and Ri, which were modeled as described below. For the same radii cited above, the mean number of slopes embraced by the influence area was 64.5 and 819."

The authors argue that each generated slope must be treated individually. Although the pluviosity events in a given time interval and inside the radius of influence of the call are similar, each slope has its own geometry and, therefore, its own values of the triad FoS/probability of failure/reliability index for both situations, saturated and natural water content. Furthermore, individual slopes may intercept different geological features at different angles. The mean number of slopes inside the influence area for each call is relatively high, and each of the risk areas presented more than one call. Therefore, the authors believe that randomly distributing these observations should preserve a similar amount of data in each slope risk area. This data, although related to the same calls and rain events over some period/location, has different input values for residence density (toe and crest), FoS/probability of failure/reliability index, distance, and angle from geological features. Additionally, the choice of the best radius now considers the smallest distance between the slope areas centroids, minimizing overlaps.

4. Is the proposed framework genuinely predictive rather than retrospective?
The rainfall accumulations are calculated relative to the emergency-call date, while the output variable aggregates calls over a time interval associated with the same event. Please specify the forecasting horizon and the exact timestamp at which a prediction would be issued. Every predictor used in an operational forecast must be available before that timestamp. A clear timeline diagram would be useful.

5. Are the evaluation metrics suitable for an early-warning application?
R^2, MAE, MSE, and RMSE are useful for regression assessment but insufficient for evaluating an early-warning system. Please report performance for confirmed mass movements separately and include event-detection metrics appropriate for decision-making, such as sensitivity, specificity, precision, recall, false-alarm rate, calibration, and lead time. Performance should also be compared against simple baseline models, such as rainfall thresholds and logistic regression.

R – In the current phase of the research, the proposed framework demonstrated a strong ability to reproduce retrospective events. However, in time series, this good performance may not be enough for predicting new events. Considering the temporally independent validation, the authors agree with the reviewer and acknowledge this limitation at the current stage of the project.

The authors, however, plan to perform temporal holdout validation during the interactive implementation of this model and gradually replace the municipality's current warning and emergency declaration criteria. We are now in the rainy period in the city. All data produced during this period will help with the model implementation task, serving as a temporal holdout validation (walk-forward validation with the use of a Sliding Windows to Test Memory Depth, since most of the variables tend to evolve, e.g., the quality of the sewage collection system, and the geometry of the slopes). At this stage, new metrics will be employed, in addition to R^2, MAE, RMSE, and MSE. Different (at least two) thresholds must be defined in this stage, considering the predicted and experimental values of the output variable: a warning and an emergency state. Declaring an emergency without confirmed events (false alarms) can jeopardize the entire system, leading the population to gradually discredit it and expose them to risk in future events, underscoring the importance of the aspects highlighted by the reviewer. However, the authors intend to present these developments in a future publication.

Please note that the output variable distinguishes between confirmed and unconfirmed calls via RFU, and that RFU values of 20 produced better results. It means that confirmed calls dominated the analysis, with a confirmed call being 20 times more important than an unconfirmed one in the calculation of the output variable.

6. How should the geotechnical stability indicators be interpreted?
The calculated Factor of Safety and probability-of-failure values are important model inputs, but they are derived from uniform-slope charts with ru = 0 and formation-level parameter distributions. Please explain whether these quantities are intended as approximate susceptibility proxies or site-specific stability estimates. The distinction is important because the conclusions currently imply a level of physical resolution that may not be supported by the assumptions.

R – The outputs of the script for slope stability calculation were compared with those of a commercial software, which produced very similar FoS and reliability index values, and a high-resolution DTM was used for slope generation. The shear strength parameters reflect field variations within the same soil formation and are the result of about two decades of soil testing at the Federal University of Bahia. However, the city morphology evolves constantly, and the performed analysis can not replace site-specific stability estimates of FoS or the reliability index. A direct investigation campaign with undisturbed sample collection and shear strength tests, along with accurate topography, among other elements, is necessary for site-specific stability estimates. This kind of study, however, is typically conducted for stabilization and containment projects, which are not the focus of this paper.

7. How will reproducibility be ensured?
The manuscript states that code and data may be available upon request and subject to institutional agreements. Sensitive emergency-call data understandably require protection. Nevertheless, reproducibility could be improved by releasing anonymized or aggregated records, the feature-construction workflow, the fitted chart coefficients, model hyperparameters, random seeds, and a synthetic demonstration dataset.

R - Please note that the paper already shares much of the information used, e.g., the quality of the sewage services, the DTM, and the presence of residences and vegetation. The FoS/probability/reliability and code can be easily shared (the authors intend to do this in the next version of the paper). However, in addition to privacy concerns about emergency-call data, the Federal University of Bahia has partnership agreements with the Salvador Municipality that prohibit data sharing unless in special circumstances. Because of that, it is postulated in the paper “... their disclosure to a third party is only possible after filtering and approval by the city administration.”

Specific comments

Lines 15-42: The literature review should more clearly distinguish among landslide susceptibility mapping, retrospective event classification, nowcasting, and operational forecasting. The novelty of the present study should be framed relative to existing municipal-scale landslide early-warning and susceptibility studies rather than primarily relative to Factor-of-Safety prediction studies.

R – The authors could not find in the literature a system similar to the one they proposed. The literature review was improved to clarify the paper's novelty and to better compare it with other studies, highlighting their differences.

Lines 27–31: Probabilistic approaches have also been applied to slope-instability assessment under uncertain loading conditions, including Bayesian-network and Newmark-displacement frameworks for earthquake-triggered landslides (Khalaj et al., 2020), as well as reliability-based methods for prioritizing maintenance actions for reinforced slopes under uncertain drainage-related failure conditions (BahooToroody et al., 2021).

R – The studies listed by the authors are indeed important contributions. However, they differ from the current study in several respects, including the output variables and the main purpose of the proposed framework. Although not completely comparable, the literature review was improved, citing related works presented previously.

Lines 75-97: Please provide the number of drained triaxial and direct-shear tests separately for each soil formation and moisture condition. Explain how results from the two test types were harmonized, what stress ranges were used, and how each slope was assigned a soil formation. The potential sampling bias arising from consultant-requested tests should also be discussed.

R – Direct shear and consolidated drained triaxial tests, when performed according to the respective standards, yield similar shear strength parameters except for highly anisotropic soils. Statistical analysis performed before the analysis revealed that the two datasets were similar; therefore, no distinction was made between the shear strength parameters from triaxial and direct shear tests. In almost all the employed tests, the stress range varied from 50kPa to 200 kPa. The consultant normally asks the laboratory to collect undisturbed samples during the site-specific investigation, and the tests are performed within the requested stress range. The authors did not realize how bias could be introduced into this process. In this study, geological maps from the city (Figure 1) were used to infer the soil formation at the middle of each slope

Lines 85-86 and Table 1: The sand-and-gravel and clay-and-silt fractions do not consistently sum to 100%. Please explain whether these values were calculated from different subsets of samples or whether a transcription error is present. Permeability values are presented only as orders of magnitude; indicate the number of measurements and the observed variability.

R – A transcription error occurred. Corrected. Thanks. The permeability dataset is limited. Therefore, the authors preferred to use typical values rather than the mean and other statistical measures. Soil water permeability is one of the parameters with the greatest variation, ranging from 1x10^-2 cm/s in gravel to 1x10^-10 cm/s in tri-laminar clays. The magnitude order is normally the parameter of concern.

Lines 100-106: Please justify the 18% slope threshold and the 5 m streamline spacing. Describe how the automated crest, toe, slope height, and slope-angle extraction procedure was validated against field observations or manually interpreted cross-sections.

R – 18%, which correspond to a slope with about 10.2 degrees of inclination is exceeded by a large margin (several standard deviations) by friction angles shown in Table 2, meaning that slopes with this inclination or less are essentially stable, even without considering cohesion and adopting saturated conditions. The obtained slopes were compared with manually computed sections after contour extraction from the DTM in QGIS, demonstrating accuracy.

Lines 107-112: The use of the Michalowski charts should be explained in greater detail. In particular, provide the fitted coefficients used in Equation 1, the interpolation procedure between slope angles, and the treatment of slopes falling outside the chart ranges.

R – The values adopted for fitting Equation 1 to the the Michalowski chart were
a = [12.4755273258759, 9.51470131119669, 8.6041342657474, 6.54667870080156, 5.50300785964871, 4.78082644886497]
b = [5.50559053664697, 5.87523752327797, 5.28983795546929, 5.06307677132193, 4.44855606473109, 3.71023586187227]
c = [4.11642414252231, 2.14208268855499, 1.29195659165675, 1.05067647046123, 0.785425316685429, 0.49248679005818]

A linear interpolations was adopted between slope angles. The Michalowski chart covers slopes with angles between 15 and 90 degrees. As no slope adopted in the analysis presented an average slope less than 15 or higher than 90 degrees, there was no point falling outside the chart ranges.

The text of the paper now include this information.

Lines 113-127: Equation 2 and the accompanying notation require revision. The symbol μ is normally used for a mean but is used here for standard deviation. The term COV is described as a coefficient of variation between variables, although the equation appears to require covariance. Please define the covariance calculation explicitly and provide the equations used to calculate probability of failure and the reliability index.

R- The reviewer is right. It is the covariance. The symbols were changed according to the reviewer's suggestion. How to calculate the probability of failure was presented in the text: "Once the values of σFoS were obtained, the probability of failure was calculated. This was the same as that for obtaining FoS < 1 (normal distribution assumed)…" The authors preserve the paper's text without the standard error and covariance equations, but they include the equation for the reliability index (βr) because it is less common.

Lines 128-136: The comparison with Slide2 is useful but insufficiently documented. Please state the total number of validation cases, the parameter combinations, the soil-profile assumptions, the groundwater assumptions, and whether identical slip-surface search settings were used. A comparison of error distributions would be more informative than mean error alone.

R – The total validation cases were 48. This information was added to the paper. The adopted parameter values fall within typical ranges in geotechnical engineering. Because the Michalowski charts are for homogeneous profiles, a single soil layer was used in the comparison. No water table was adopted for ru=0. In the case of the Michalowski charts, no slip-surface search is adopted. In Slide2, appropriate entry and exit ranges were adopted, ensuring the critical slip surface was not located at any of the entry/exit ranges' extremes.

Lines 137-139: The manuscript states that most emergency calls are concentrated in areas with low saturated Factor of Safety. Please quantify this relationship and report the proportions of confirmed and unconfirmed calls across Factor-of-Safety classes.

R – Unfortunately, because many slopes are considered in a single call, whether confirmed or not, each with its own FoS, the authors could not fulfill the reviewer's request.

Lines 145-152: The representation of geological structures requires additional explanation. Please justify the use of the sine of the angle between the slope section and the nearest structure. Clarify

whether structure dip, distance, persistence, and multiple nearby structures were considered or omitted.

R – The worst scenario occurs when a geological structure intersects the slope direction perpendicularly, having a negligible effect when slope direction and geological structures are parallel. This variable can be considered the complement of the scalar product between the vectors; therefore, the sine of the angles was used in this representation. Only the nearest structure is considered. Structures outside the radius of influence of the slope are not considered. The dips of the faults were not considered because not all faults have this information, and the dip varies with depth at the same point. Since the geological and geotechnical scales are very different, considering dip at this moment in the paper was not possible. This information was added to the text.

Lines 153-158: The statement that residences near the slope toe tend to have a stabilizing effect requires qualification and supporting evidence. Buildings may introduce toe loading, but construction may also involve excavation, drainage modification, and additional water infiltration. Please avoid assigning a universally stabilizing effect without analysis.

R – Buildings at the slope toe have a stabilizing effect because they create an effect similar to a berm of soil. Of course, performing excavation before construction (which is not usual here due to climatic conditions) will reduce this beneficial effect. The opposite occurs at the slope crest, where the slip surfaces are steeper near the soil surface and almost all the loading is converted in shear stress. Please note that the text refers solely to the effect of the density of residences. The authors are aware that there are other destabilizing effects in a risk area. The increase in infiltration rates, for instance, was indirectly considered in the paper in terms of the quality of sewage services.

Lines 159-162: Please describe rainfall-data completeness, station coverage over time, treatment of missing records, and validation of the inverse-distance-squared interpolation. It would be useful to quantify interpolation uncertainty, especially in areas distant from weather stations.

R – The paper text was modified to “The weather stations presented in Figure 4 provided hourly information on rainfall in the period of analysis (2020-2025). For each eligible slope, the accumulated rainfall for 1d, 2d, 3d, 5d, 7d, 10d, 15d, and 30d was calculated using as a reference the date of the emergency call and the values recorded in the weather stations, which were interpolated using the inverse of the square distance as weight. All the weather stations provided valid rainfall records on almost all days. In the case of weather stations with no valid data on a given day, the interpolation procedure was performed using the remaining weather stations”

Unfortunately, the authors do not have any independent pluviometers available to quantify interpolation uncertainty in regions between the weather stations.

Lines 163-177: The sewage-service score is potentially informative but the assigned weights require justification. Please explain the basis for assigning values of 0, 0.1, 0.2, 0.3, 0.6, and 1.0 and provide a sensitivity analysis. The variable may also act as a proxy for socioeconomic conditions, reporting behaviour, and infrastructure quality; this should be acknowledged.

R – These values were assumed first by intuition and then trying to get a better positioning of this input variable in terms of gain and weight in the obtained results. The text already discuss the nature of the variable “This input variable reflects the development status of the population in the census sector and, to some extent, the presence of the public services. Furthermore, areas with no sewage collection systems (and probably no proper drainage systems) suffer from what may be

called anthropic rainfall, as practically all the sewage water infiltrates the terrain along with part of the rainfall.”

Lines 178-181: Please describe how stabilization measures were mapped and dated. Because interventions are likely to target slopes already considered hazardous, this variable may be affected by confounding by indication. Its interpretation should therefore be cautious.

R – Stabilization areas were informed by municipality with the respective construction period, location and extension. Very few slopes were located in areas that suffered intervention in the period of analysis, unfortunately:

Lines 202-204: Replace “based on threes” with “based on trees.”

R – Done. Thanks

Lines 205-209: The discussion of artificial neural networks is not relevant to the adopted XGBoost method and should be removed or rewritten to address gradient-boosting hyperparameter optimization specifically.

R – As the paper may have a very varied audience, with different knowledge levels about ML, the authors decide to preserve the text in the paper.

Lines 210-213: Please report the RandomizedSearchCV parameter ranges, number of sampled combinations, cross-validation procedure, scoring function, and random seed.

R - The hyperparameter optimization of the XGBoost model was done using RandomizedSearchCV. The parameter ranges used were: number of boosting trees used [100, 2500], overly large values can cause overfitting; learning_rate [0.01, 0.5], controls the learning speed. Training is faster with a high value, but might overshoot optimal solutions; maximum depth [3, 15], controls model complexity and training speed; subsample [0.1, 0.9], used to build each tree, helping reduce overfitting; colsample_bytree [0.1, 0.9], fraction of predictor variables randomly selected for each tree, to reduce overfitting; L1 regularization term [0, 5], to add sparsity and reduce overfitting; L2 regularization term [0.01, 5]; to improve stability and reduce overfitting; minimum sum of instance weights [1, 10], required in a child node, limiting overly specific tree splits; maximum number of bins [256, 2048], used in the tree splits.

The final optimized model adopted the following hyperparameter values: n_estimators = 2310, learning_rate = 0.1186, max_depth = 13, subsample = 0.9674, colsample_bytree = 0.9552, reg_alpha = 3.6742, reg_lambda = 1.5491, min_child_weight = 3, and max_bin = 1353. The XGBoost regressor was configured with the gbtree booster, reg objective function, and hist tree construction method.

This information was added to the paper

Lines 218-228: A radius of influence of 300 m produces the highest R^2, whereas 150 m is selected as the preferred value based on physical consistency. Please define an objective selection criterion and distinguish between predictive performance and spatial interpretability.

R – Paper now uses the distance between the centroids of the nearest areas to choose, in a more objective way, the best radius value in the performed analysis.

Lines 230-253: Feature-importance interpretation should be strengthened. Several predictors are strongly related or mathematically dependent, including rainfall accumulations over nested durations and the Factor-of-Safety, probability-of-failure, and reliability-index variables. Split weight and gain may therefore be misleading when interpreted individually. Please consider permutation importance or SHAP analysis and avoid causal interpretations.

R – The authors will improve feature-importance interpretation by adding a SHAP analysis of the best radius obtained, considering geological morphology.

Lines 255-280: The conclusions substantially overstate the demonstrated capability of the current model. Claims regarding accurate real-time prediction, evacuation warnings, critical thresholds, scalability, and proactive mitigation should be tempered until the model has been evaluated using a leakage-resistant spatial and temporal holdout design and tested prospectively.

R – The conclusions now cite the need of evaluation of the model with temporal holdout validation during its implementation

“Although in its present form, the proposed framework demonstrated a strong ability to reproduce retrospective events in time series, this good performance is usually not sufficient to ensure strong predictive performance. Temporal holdout validation and an interactive implementation of this model are under development, aiming to gradually replace the municipality's current warning and emergency declaration criteria. All the new data produced during this period will help with the model implementation task, using walk-forward validation with Sliding Windows to Test Memory Depth, since most variables tend to evolve (e.g., the quality of the sewage collection system and the geometry of the slopes). Different (at least two) thresholds must be defined in this stage, considering the predicted and experimental values of the output variable: a warning and an emergency state. This is a very delicate phase, because declaring an emergency without confirmed events (false alarms) can jeopardize the entire system, leading the population to gradually discredit it and expose them to risk in future events”

Lines 283-288: The code-availability, data-availability, and combined code-and-data-availability statements are repetitive. Please consolidate them into one clear statement and describe which anonymized materials can be released.

R – Corrected. Thanks

Lines 342-343: The DOI provided for Morgenstern and Price (1965) is incorrect. Please replace it with: 10.1680/geot.1965.15.1.79.

R – Corrected, thanks

Lines 351-352: The Pereira and Barbosa (2023) reference is incorrect and contains a malformed DOI. The reference should identify the censobr R package: “censobr: Download Data from Brazil’s Population Census,” with DOI: 10.32614/CRAN.package.censobr.

R – Corrected, thanks

Lines 368-369: This reference conflates ter Braak and Šmilauer with the R Core Team citation.

If the authors intend to cite the R environment, the reference should be attributed to the R Core Team. If the Canoco manual is not used in the study, the ter Braak and Šmilauer reference should be removed.

R – Corrected, thanks

Technical corrections

Line 13: Capitalize “Keywords.”
Line 21: Remove the unmatched closing parenthesis after unit weight γ.
Line 61: Add a comma after “occur.”
Line 75: Use consistent ASTM standard designations, including the applicable version suffixes.
Line 87: Replace “Atteberg” with “Atterberg.”
Lines 96-97 and Table 2: Replace “unity weight” with “unit weight” and correct the units to kN/m³.
Line 102: Correct the DTM dimensions. The current formatting “5,4299 × 5,3010 pixels” appears erroneous.
Line 180: Replace “emergence call date” with “emergency-call date.”
Line 202: Replace “threes” with “trees.”
Figure 6 caption: Replace panel label “w)” with “e).”
Table 3: Remove the duplicated “20” in the row for Ri = 150 m.
Line 243: Replace “second platoon” with “second group” or “second tier.”
Line 247: Replace “2th place” with “2nd place.”
Line 272: Replace “phisically” with “physically.”
Line 290: The initials “SLM” are repeated in the formal-analysis and validation contribution statement.
Competing-interests statement: Replace “The authors declare have no competing interests” with “The authors declare that they have no competing interests.”
Line 364: Correct the spelling of “Salvador” in the Souza-Oliveira et al. reference.
Review the manuscript throughout for spacing between numerical values and units, including “150 m,” “4 d,” and “75 m.”

R – All done. Thank you very much. The authors really appreciate all your throughout revision,

.