



Modelling the rainfall threshold for shallow landslides considering the landslide predisposing factors in Portugal

Abstract Rainfall-triggered landslides pose a significant threat to both infrastructure and human lives, making it crucial to comprehend the factors that contribute to their occurrence. Specifically, understanding the relationship between these factors and the amount of rain that is necessary for triggering such events is essential for effective prediction and mitigation strategies. To address this issue, our study proposes a statistical modelling approach using machine learning, specifically the Random Forest algorithm, to investigate the connection between landslide predisposing factors and the daily rainfall intensity threshold necessary for the initiation of shallow landslides in Portugal. By leveraging a comprehensive dataset comprising historical landslide events, associated critical rainfall, and ten distinct landslide predisposing factors, we developed several models and used cross-validation technique to evaluate their performance. Our findings demonstrate that the Random Forest model effectively captures a relationship among landslide predisposing factors, critical daily rainfall intensity, and landslide occurrences. The models exhibit a satisfactory accuracy in assessing the spatial variation of critical daily rainfall intensity based on the predisposing factors, with a mean absolute percentage error (MAPE) of around 17%. Furthermore, the models provide valuable insights into the relative importance of various predisposing factors in landslide triggering, highlighting the significance of each factor. It was found that it takes higher rainfall intensity to trigger shallow landslides in the north region of Portugal when considering critical rainfall events of 3 and 13 days. Slope aspect, slope angle, and clay content in the soil are among the main predisposing factors used for defining the spatial variation of the daily rainfall intensity threshold.

Keywords Critical rainfall intensity · Predisposing factors · Shallow landslide · Random Forest

Introduction

Landslides are a common natural hazard that have caused heavy damage all over the world, including in Portugal (Campbell 1975; Guzzetti et al. 2007; Zêzere et al. 2015). There is a great variety of landslides due to the diversity of conditions that cause slope instability and the dynamic processes that trigger the events (Popescu 2002). Shallow landslides in specific are characterised by small thickness (few dozen centimetres to a few metres) and are usually triggered by medium to very intense rainfall events (Montrasio and Valentino 2008; Zêzere et al. 2005). The factors that control the occurrence of shallow landslides can be divided into the quasi-static variables and the dynamic variables (Wu and Sidle 1995). The quasi-static variables (e.g. soil properties, topography, and geology)

define the spatial distribution of the landslides, while the dynamic variables (e.g. soil saturation, climatic conditions, hydrological processes, and human activities) control the time of occurrence of landslides.

Throughout the previous decades, landslide susceptibility modelling has attracted the attention of various researchers around the world (e.g. Chen et al. 2018; Reichenbach et al. 2018; Zêzere et al. 2017); however, landslides still are a global danger and it is still a challenge to predict where and when they will occur. Statistical techniques are a popular method for the creation of landslide susceptibility maps. Within this approach, the researchers must consider the type of landslides, analysis scale, study area characteristics, and availability of datasets, and then proceed with the selection of method and landslide predisposing factors for their models (Zêzere et al. 2017). Recent studies have been including the analysis of dynamic variables influencing landslide occurrence in order to create spatial-temporal forecasting models for landslides (Tehrani et al. 2022; Nocentini et al. 2023; Fang et al. 2024).

In recent years, statistical modelling approaches employing machine learning techniques have gained popularity in landslide research, as they offer the potential to uncover complex relationships between multiple variables (Merghadi et al. 2020). The machine learning techniques have been generally enriching the quality and accuracy of generated landslide susceptibility maps and were confirmed to achieve improved performance when compared to classical methods for defining landslide rainfall thresholds (Chen et al. 2019; Distefano et al. 2022).

Regional variations of rainfall intensity-duration thresholds for landslide triggering have been identified in many different works, suggesting that some environmental factors may be responsible for the threshold definitions (Wieczorek and Guzzetti 1999). Although many studies have been published about landslide susceptibility and rainfall thresholds, it is not common to find works that relate predisposing factors with the critical rainfall event that triggered the landslides (Palladino et al. 2018; Peruccacci et al. 2017). Understanding how the triggers of landslides react with different predisposing factors is critical for effective landslide prediction and mitigation strategies. The identification of predisposing factors and their relationship with the critical rainfall event can provide valuable insights into the mechanisms underlying landslide occurrence, thus increasing the accuracy of landslide forecasting.

In this study, we focus on the creation of statistical models to investigate the relationship between the predisposing factors and the amount of rainfall necessary to trigger shallow landslides in Portugal. We used a comprehensive dataset that includes historical landslide events, as well as the associated critical rainfall and

predisposing factors. The data were collected from different sources and the predisposing factors considered include elevation, slope angle, slope aspect, lithology, distance to rivers and faults, and some soil properties (water content at field capacity, clay content, coarse sand, and coarse grains in the soil). The Random Forest (RF) algorithm, a machine learning technique capable of capturing nonlinear relationships and handling large datasets (Taalab et al. 2018), was employed to create the models and assess the correlation of the shallow landslide predisposing factors and critical daily rainfall intensity. The final goal is to develop an analysis of the spatial variation of the daily rainfall intensity thresholds considering the shallow landslide predisposing factors for all of mainland Portugal.

Methodology

The study area

The area for this research is the continental region of Portugal (Fig. 1a). Portugal is a country located in southwestern Europe, on the Iberian Peninsula, and it is bordered by the Atlantic Ocean to the west and south, and by Spain to the north and east. Portugal has a diverse landscape that includes coastal plains, mountains, and plateaus. The country can be broadly divided into three regions: the mainland, the Azores, and Madeira. Our work will focus on the mainland area that is characterised by its varied topography, with the Tagus River being the longest river in the country, and the Serra da Estrela being the highest mountain range (1993 m asl). In terms of size, Portugal has a total land area of approximately 92,000 km², making it one of the smaller countries in Europe. The mainland constitutes the largest portion of the country's land area.

The climate of Portugal is influenced by its latitude and relative position to the Atlantic Ocean and is generally classified as Mediterranean. The summer is hot and dry, especially in the interior regions, with average temperatures ranging from 25 to 30 °C. The winter is mild and rainy, with average temperatures ranging from 8 to 15 °C (Vaz 2021). However, there are regional variations in climate due to the country's diverse geography. The coastal areas tend to be milder and more humid, while the interior regions are much hotter and drier. The northern and southern regions experience a notable temperature contrast throughout the year. Rainfall in Portugal is highest during the winter months, with the northwest region receiving the most precipitation (Fig. 1b). The southern regions, such as the Algarve, are generally drier and hotter. Portugal also experiences occasional extreme weather events, such as heatwaves and extreme rainfall events generating floods and landslides.

Data

The landslide inventory used in this work was extracted from two databases that were built based on information extracted from newspapers: (1) the DISASTER database that only considers events that caused any disturbance in people's lives (Zêzere et al. 2014); (2) a second database of landslides that did not cause damage to humans (Vaz 2021). Both databases were merged into one to make the statistical analysis. The unified database includes different types of landslides and contains the information of the day and the coordinates of occurrence. The spatial accuracy of the landslide coordinates is classified in 5 classes (Zêzere et al. 2014): C1 (exact coordinates, scale 1:1000), C2 (based on local toponymy, scale 1:10,000), C3 (based on local geomorphology, scale 1:25,000), C4

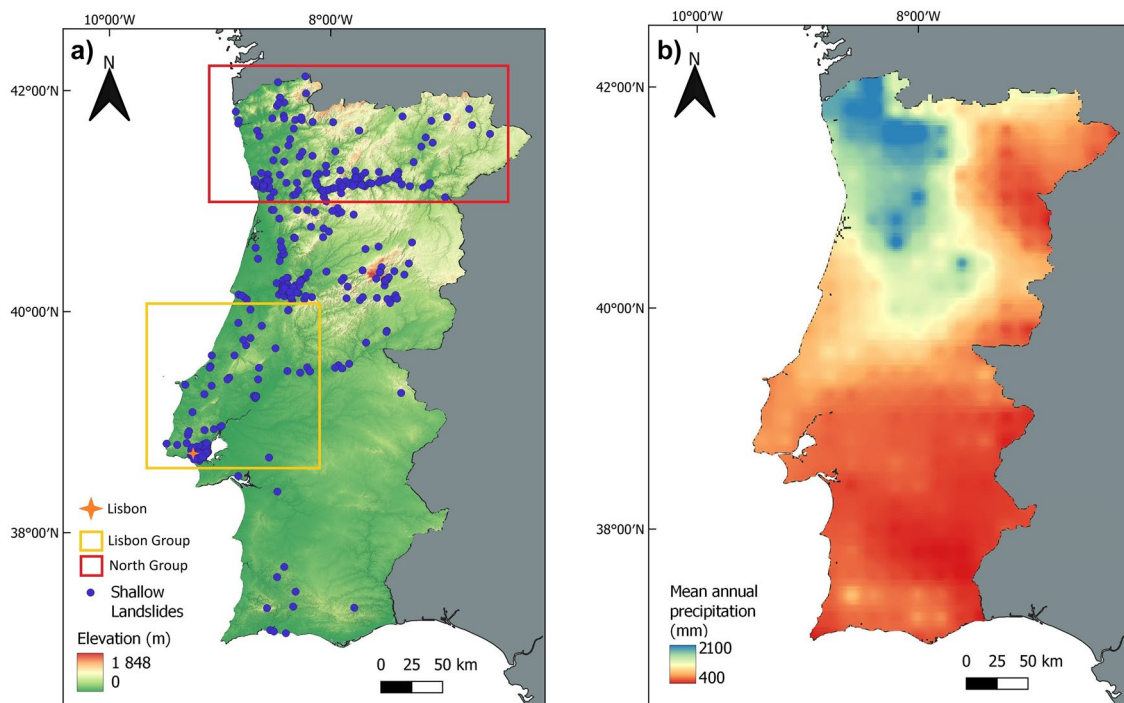


Fig. 1 Shallow landslides inventory distributed throughout mainland Portugal (a) and mean annual precipitation (1950–2003) (b)

(centroid of the parish), and C5 (centroid of the municipality). Only the first three classes were used in this work.

The final database was filtered to retain only the landslide events that occurred after the year of 1950, in order to agree with the temporal range of some of the other data sources used in this work. Any occurrence identified as non-slide type was eliminated. The landslides extracted from the database had to be further filtered into shallow landslides. For the identification of shallow landslides, we have automated in Python a methodology (Fig. 2) to extract and process the daily rainfall data from 1950 to 2022, downloaded from the Copernicus database with a resolution that ranges from $0.1^\circ \times 0.1^\circ$ and $0.25^\circ \times 0.25^\circ$ (Cornes et al. 2018).

Figure 2 shows the diagram of the method that consists in extracting the critical rainfall event duration and critical daily intensity that triggered the landslide events to classify them into categories of “most likely to be a shallow landslide” and “unlikely to be a shallow landslide”. The main idea of the method is that rainfall-triggered shallow landslides in Portugal have been normally triggered by precipitation lasting from 1 to 15 days (Zêzere et al. 2005). In order to extract the critical rainfall event duration that triggered the landslides, we used a similar method from Brunetti et al. (2010), which considered that the minimum dry interval that separates two rainfall events is 48 h for the dry seasons and 96 h for the wet seasons. The dry period in Portugal is between the months of May and September and the wet period between October and April (Vaz 2021). Similarly to Vaz (2021), a daily precipitation

threshold of 1 mm was considered, below which the precipitation is classified as insignificant. Finally, to consider precipitation as the triggering factor of the landslides, it was necessary to detect precipitation higher than 20 mm in the 5 days prior to the slope instability event (Valenzuela 2017).

It was also calculated the critical cumulated rainfall divided by the critical rainfall duration preceding each of the shallow landslide events, resulting in the critical daily rainfall intensity. The critical daily rainfall intensity was used as the triggering factor (dependent variable) in our models. Figure 3 shows the histogram of the filtered shallow landslide events and their associated critical daily rainfall intensity, and Fig. 4 shows the cumulated rainfall for each critical rainfall duration. Due to the imbalance nature of the rainfall intensity dataset shown on the histogram (Fig. 3), where values between 10 and 20 mm/day concentrate most of the data, only the range with more samples of rainfall intensity was used as the dependent variable for all models to acquire better performance (see “Models”), resulting in a final dataset of 186 samples.

The landslide predisposing factors considered in this study (Fig. 5) include elevation, slope angle, slope aspect, lithology, distance to rivers, distance to faults, and some soil properties (water content at field capacity, clay content, coarse sand, and coarse grains in the soil). Table 1 summarises the landslide predisposing factors and the respective range of values.

Elevation can often define different climatic and vegetation zones, while slope angle directly controls the shear forces acting

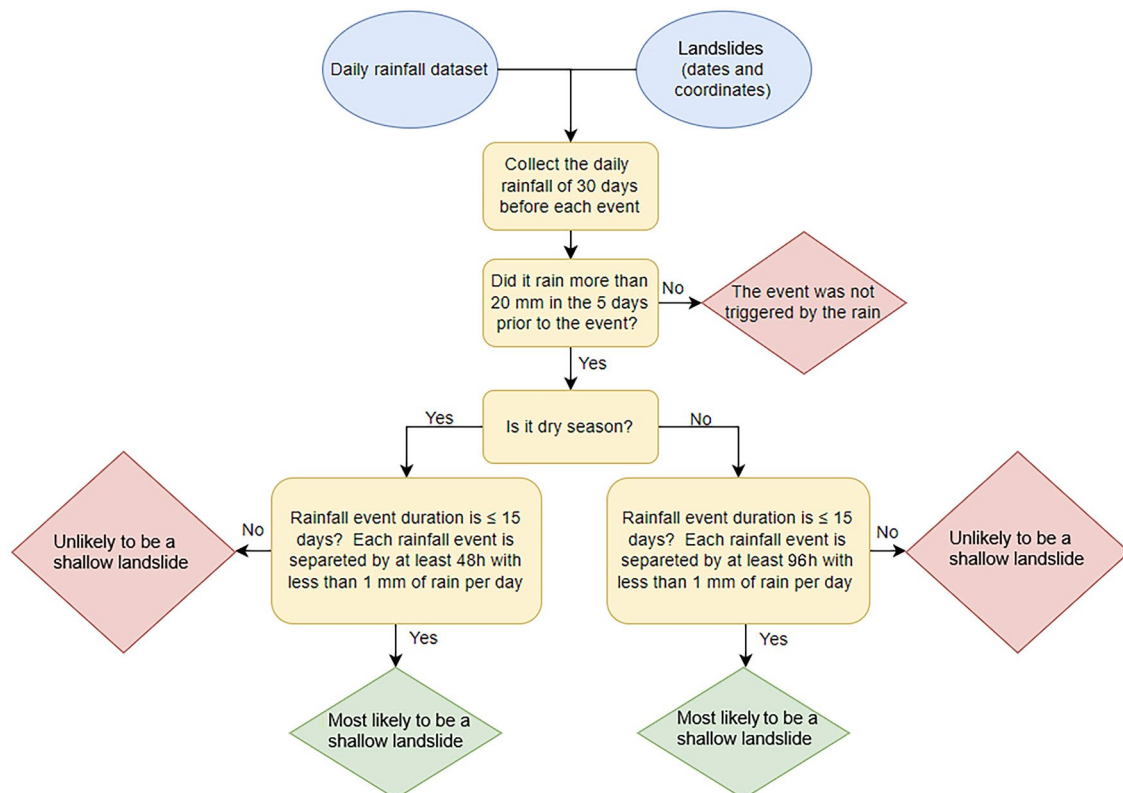


Fig. 2 Flowchart of the methodology to classify the landslides into categories of “most likely to be a shallow landslide” and “unlikely to be a shallow landslide” taking into account the rainfall prior to the event

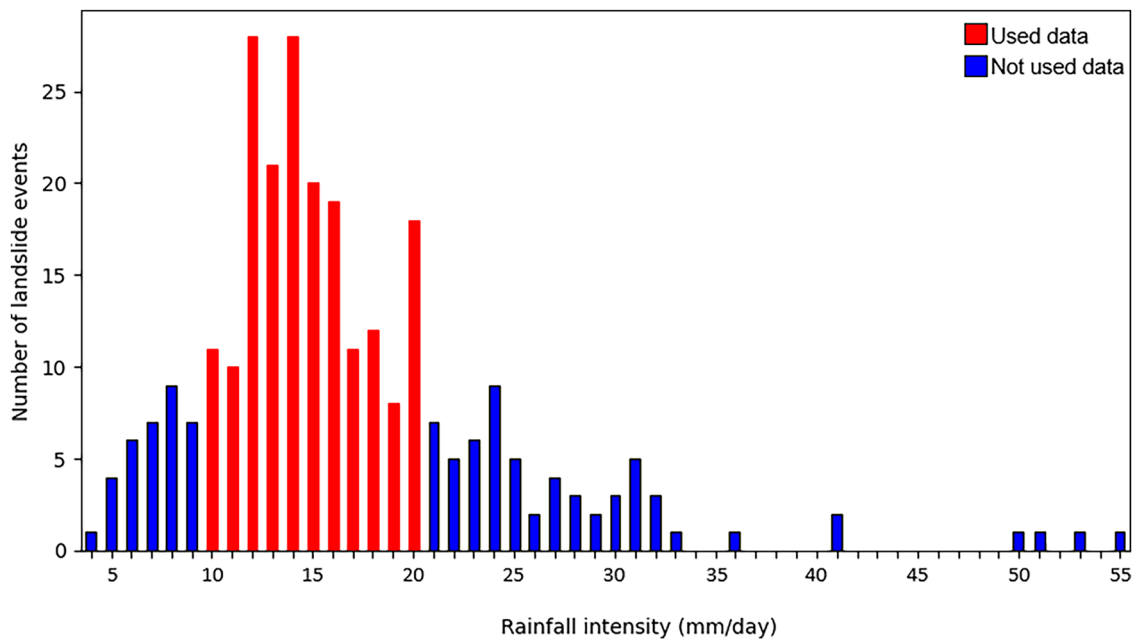


Fig. 3 Histogram of the range of values of critical daily rainfall intensities of the shallow landslides (in red are the data used in the models)

on hillslopes. The slope aspect can influence the amount and intensity of rainfall that affects the hillslope, as well as the amount and intensity of solar radiation. It was found that the

slope aspect is an important predisposing factor when dealing with superficial landslides, namely in clayey soils (Capitani et al. 2013). Lithology represents the properties of bedrock and

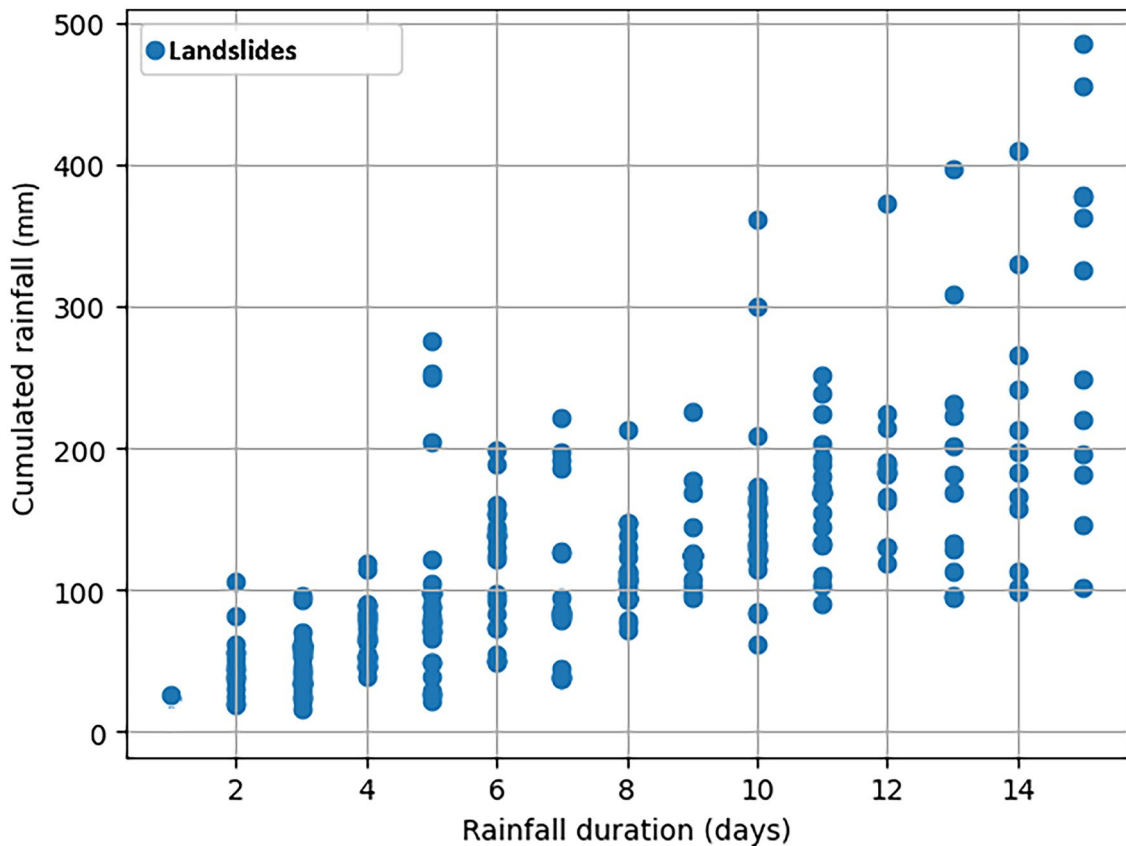


Fig. 4 Cumulated critical rainfall vs critical rainfall duration for each shallow landslide occurrence

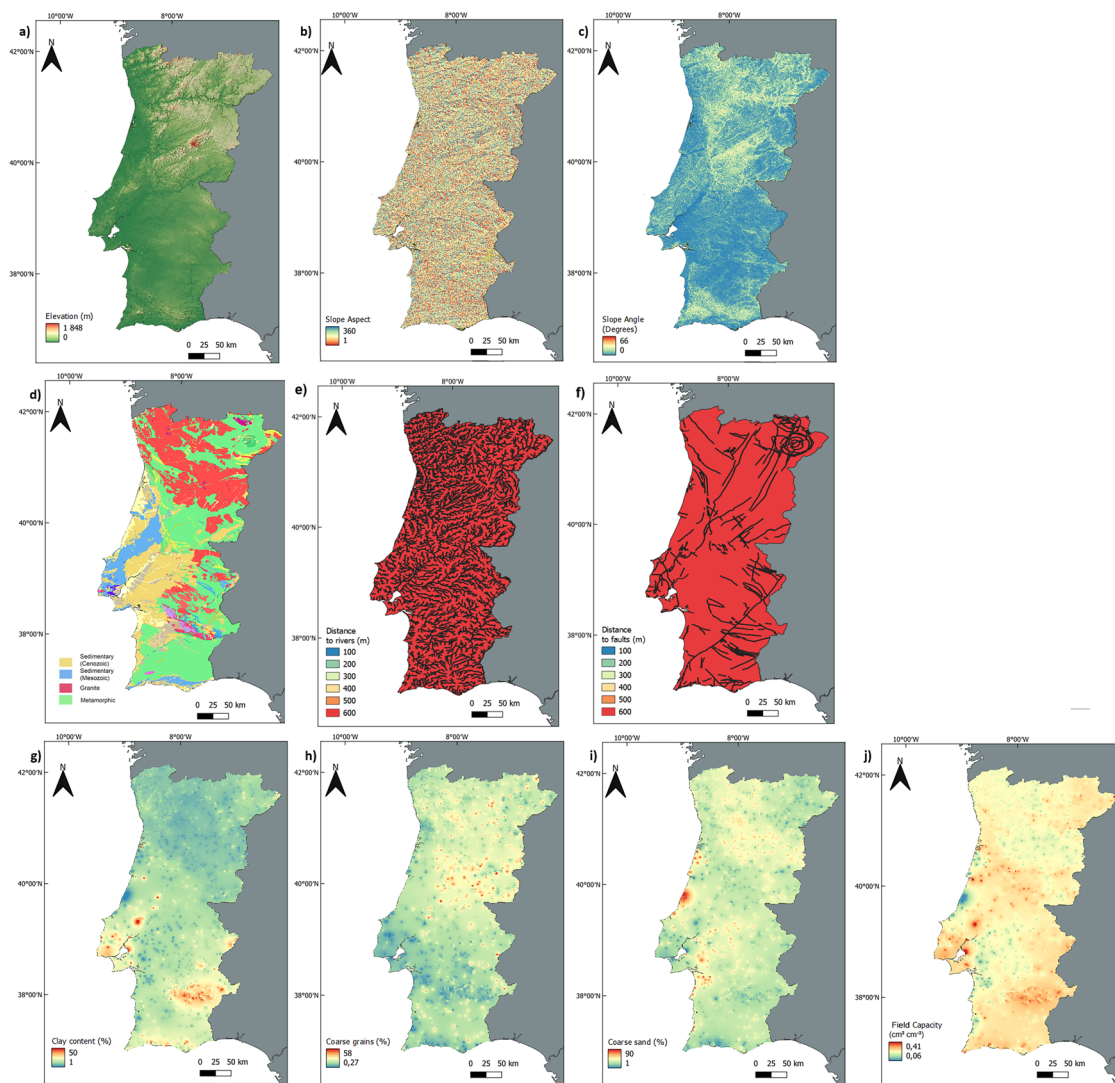


Fig. 5 Map of the 10 landslide predisposing factors used in this work: **a** elevation; **b** slope aspect; **c** slope angle; **d** lithology; **e** distance to rivers; **f** distance to faults; **g** clay content; **h** coarse grains; **i** coarse sand; **j** field capacity

the soil properties influence cohesion, water infiltration, and percolation. Distance to rivers could indicate the proximity to groundwater and detect areas more prone to erosion (Gómez and Kavzoglu 2005).

Similar predisposing factors have been largely used in various studies related to susceptibility modelling and have shown to be efficient for landslide prediction (e.g. Chen et al. 2019; Nsengiyumva and Valentino 2020; Catani et al. 2013; Lima et al. 2022). We assume that those same predisposing factors should also have a relation in defining the critical rainfall intensity necessary to cause the landslides.

The morphological factors were extracted from a DEM with 30 m/pixel resolution. Lithology and distance to faults were collected from the Portuguese Geological Survey (LNEG). Distance to rivers was extracted from a hydrographic network provided by the Agência Portuguesa do Ambiente (APA). The soil characteristics were collected from the Infosolos database (Ramos et al. 2017).

Models

The main goal of this research is to identify the spatial variation of shallow landslide daily rainfall intensity threshold based on the landslide predisposing factors. For this reason, we elaborated the models using the RF algorithm, a popular machine-learning technique (Breiman 2001; Vorpahl et al. 2012; Catani et al. 2013). The RF was chosen because of its capabilities of analysing nonlinear relations between a complexity and variety of independent variables, and it outputs the score (importance) of the independent variables used in the predictions (Taalab et al. 2018). RF has also shown excellent results in previous landslide studies, especially those assessing landslide susceptibility (Pourghasemi et al. 2020; Chen et al. 2017; Liu et al. 2021; Rosi et al. 2023).

We partitioned the model development into two phases, the initial phase, and the final phase. During these phases, different subsets of events were extracted from the total 186 filtered occurrences

Table 1 Predisposing factors and respective range of values, source, and scale/resolution

Predisposing factor	Range of values	Source	Scale/ resolution
Distance to rivers	0-600 m (groups of 100)	APA	1:100.000
Distance to faults	0-600 m (groups of 100)	LNEG	1:500.000
Lithology	Categorical	LNEG	1:1.000.000
Slope angle	0-66 degrees	Copernicus DEM	30 m/px
Slope aspect	0-360 degrees	Copernicus DEM	30 m/px
Elevation	0-2000 m	Copernicus DEM	30 m/px
Clay content	0-50%	Infosolos	1000 m/px
Field capacity	0.05-0.4 (cm ³ cm ⁻³)	Infosolos	1000 m/px
Coarse grains	0-65%	Infosolos	1000 m/px
Coarse sand	0-96%	Infosolos	1000 m/px

(see “Data”) to elaborate the models. In the initial phase, we created four distinct testing models to verify the potential of the methodology for extracting insights about the correlation between critical daily rainfall intensities and landslide predisposing factors (Table 2). During the final phase, we created three final models (Table 2) to assess the spatial variation of the shallow landslide daily rainfall intensity threshold in Portugal.

The database was split into training set (70%) and testing set (30%) for all the models elaborated in this work. Cross-validation can be especially useful when the amount of data is limited like in our case, since it uses all the training dataset, avoiding data loss. K-fold cross-validation was performed across 10 folds on the training set of testing model 1, exploring various numbers of trees to determine the optimal parameter value. The number of trees was defined as 500 since it is the value where the errors start stabilising (Fig. 6). The optimal number of trees found for testing model 1 was repeated in the other models. To assess the performance of the models, we applied root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) for all regression models and the metrics accuracy and precision for the classifier model.

The initial phase was conducted after analysing the spatial distribution of the critical daily rainfall intensities considered to have triggered the shallow landslides in the database (shown in “Rainfall intensity distribution”, Fig. 7). We analysed that the distribution of the critical daily rainfall intensity highly matches the distribution of some of the landslide predisposing factors, especially clay content. Thus, it was important to develop four testing models to improve the confidentiality of this method’s ability to reveal the correlation between predisposing factors and critical daily rainfall intensity, ensuring it goes beyond merely examining Portugal’s geographical aspects.

The testing model 1 was created to correlate the landslide predisposing factors with the corresponding critical daily rainfall intensity. The data used consists of all 186 occurrences present in the final dataset, which was divided into independent variables (landslide predisposing factors) and the dependent variable (critical daily rainfall intensity).

The other three testing models were created to compare how they score the landslide predisposing factors in relation to testing model 1. We divided the landslide dataset into two geographical groups that were delimited based on the map of clay content in the soil, where one group (North group) concentrates on 40 landslide events located in regions with low percentage of clay in the soil, while the other group (Lisbon group) contains 65 landslides that occurred in a region with high clay content (Fig. 1a).

Testing model 2 is a binary classifier, so the dependent variable contains two classes (North group and Lisbon group). The aim of testing model 2 is to understand how RF would differentiate the two classes in terms of the predisposing factors. North group and Lisbon group were analysed separately by testing models 3 and 4 respectively. The goal of testing models 3 and 4 is to verify how the same algorithm used to correlate the landslide predisposing factors with the corresponding critical daily rainfall intensity in testing model 1 would perform in a restricted area with less regional variations of predisposing factors and mean annual precipitation.

The final phase aims at the elaboration of interpretable maps of the spatial variations of rainfall thresholds for the study area. To correctly interpret the maps, it was essential to set a specific duration for the critical rainfall and observe how the rainfall intensity threshold varies, focusing on just one duration value. The problem of filtering only one value of critical rainfall duration is that we would end up with not enough data available for creating a trustful model. On the other hand, if we considered a fixed critical rainfall duration for all the landslides in the dataset, we would overvalue or undervalue the real critical rainfall intensities. For this reason, we created three groups to restrict the range of critical rainfall durations and the fixed critical rainfall duration is the median value of the durations of each group. The first group only considered shallow landslides with critical rainfall durations ranging between 1 and 5 days (fixed duration of 3 days), which represents 65 events out of the total 186 occurrences. The second group contains shallow landslides with critical rainfall duration ranging between 6 and 10 (fixed duration of 8 days), which represents 74 samples, and the third group considers shallow landslides with critical rainfall

Table 2 Description of the models developed in this work

Model	RF variables	Data used	Goal	No. of samples
Testing model 1 (regression)	Dependent: critical daily rainfall intensity Independent: predisposing factors	All shallow landslides that were filtered for mainland Portugal	Relate the predisposing factors with the corresponding critical rainfall intensities taking in account the whole dataset	186
Testing model 2 (classifier)	Dependent: North group and Lisbon group (Fig. 1) Independent: predisposing factors	Dataset segregated into North group and Lisbon group (Fig. 1)	Understand how the model would rank the predisposing factors to differentiate North region and Lisbon region	105
Testing model 3 (regression)	Dependent: critical daily rainfall intensity Independent: predisposing factors	Only shallow landslide in the North group	Understand how the model would rank the predisposing factors within a restricted area with less variation of the predisposing factor values	40
Testing model 4 (regression)	Dependent: critical daily rainfall intensity Independent: predisposing factors	Only shallow landslide in the Lisbon group	Understand how the model would rank the predisposing factors within a restricted area with less variation of the predisposing factor values	65
Final model 1 (regression)	Dependent: critical daily rainfall intensity (3 days of duration) Independent: predisposing factors	Shallow landslides triggered by critical rainfall duration ranging from 1 to 5 days	Analyse the spatial variation of the daily rainfall intensity threshold	65
Final model 2 (regression)	Dependent: critical daily rainfall intensity (8 days of duration) Independent: predisposing factors	Shallow landslides triggered by critical rainfall duration ranging from 6 to 10 days	Analyse the spatial variation of the daily rainfall intensity threshold	74
Final model 3 (regression)	Dependent: critical daily rainfall intensity (13 days of duration) Independent: predisposing factors	Shallow landslides triggered by critical rainfall duration ranging from 11 to 15 days	Analyse the spatial variation of the daily rainfall intensity threshold	67

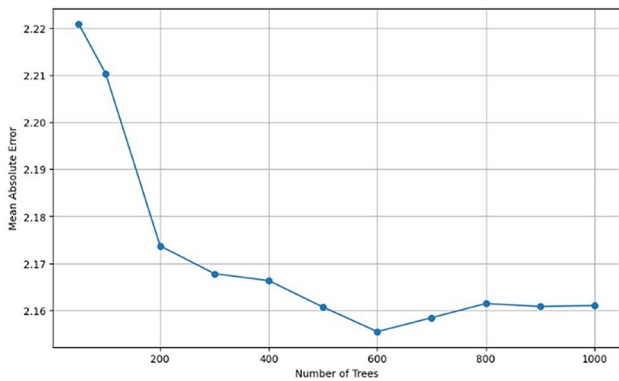


Fig. 6 RF model error in relation to the increase in the number of trees

duration ranging between 11 and 15 days (fixed duration of 13 days), totalling 67 samples. With this approach, we aim at minimising the effects of overvaluing and undervaluing the real critical rainfall intensities. Three models were created for each of the three groups, using landslide predisposing factors as independent variables and the critical daily rainfall intensity as the dependent variable.

Results

Rainfall intensity distribution

As a first step to understand the relationship between the predisposing factors and the triggering factor (critical daily rainfall intensity) that caused shallow landslides in mainland Portugal, we have plotted the landslide occurrences with the respective critical daily rainfall intensity related to each event. Figure 7 shows an interpolation of the plot of the critical daily rainfall intensity that triggered each occurrence of shallow landslides for the three groups created during the final phase. The maps highlight a regional difference in terms of critical daily rainfall intensity distribution. The northern region has a concentration of higher critical daily rainfall intensity in comparison with the Lisbon region. These regional differences

are in conformity with previous works such as Vaz (2021), which shows that northern areas have higher rainfall thresholds for triggering landslides compared to the Lisbon area.

Influence of predisposing factors

Initial phase

Figure 8 shows the predisposing factor importance for the four testing models created using the RF algorithm. During landslide predisposing factor selection and to acquire the best performance, the predisposing factor distance to faults was eliminated from the four models due to its low importance, meaning that it was working as noise and had almost no positive correlation with the model prediction capability.

After predisposing factor selection, testing model 1 shows that the first two predisposing factors that most improved the accuracy of the model were the clay content in the soil and the slope aspect. The factors distance to rivers and lithology were the two least scored predisposing factors meaning low correlation with the dependent variable.

Testing model 2 was built to check how the predisposing factors can differentiate two different landslide-prone regions in Portugal and how it ranks the landslide predisposing factors' importance differently from testing model 1. Elevation, slope aspect, slope angle, and distance to rivers have almost no importance for the second model, while the clay content in the soil is the most important one by a large difference compared to the second-ranked predisposing factor (field capacity). Testing model 2 shows significant differences and similarities in comparison with testing model 1 in terms of predisposing factor importance. First, the clay content in the soil is represented as a very important predisposing factor for both testing models 1 and 2, even though in testing model 2 the clay content represents a more relevant part of the model importance (> 40%). On the other hand, the slope aspect, elevation, and slope angle had almost no importance in testing model 2, while they were the first, second, and third most important predisposing factors in testing model 1.

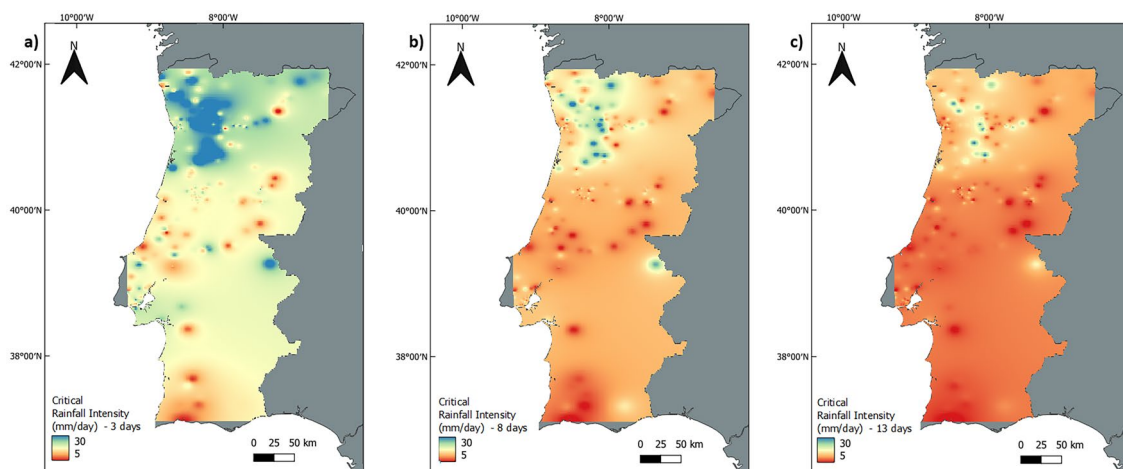


Fig. 7 Interpolation of rainfall intensity values for the three groups of critical rainfall duration (a 3 days of critical rainfall duration; b 8 days of critical rainfall duration; c 13 days of critical rainfall duration)

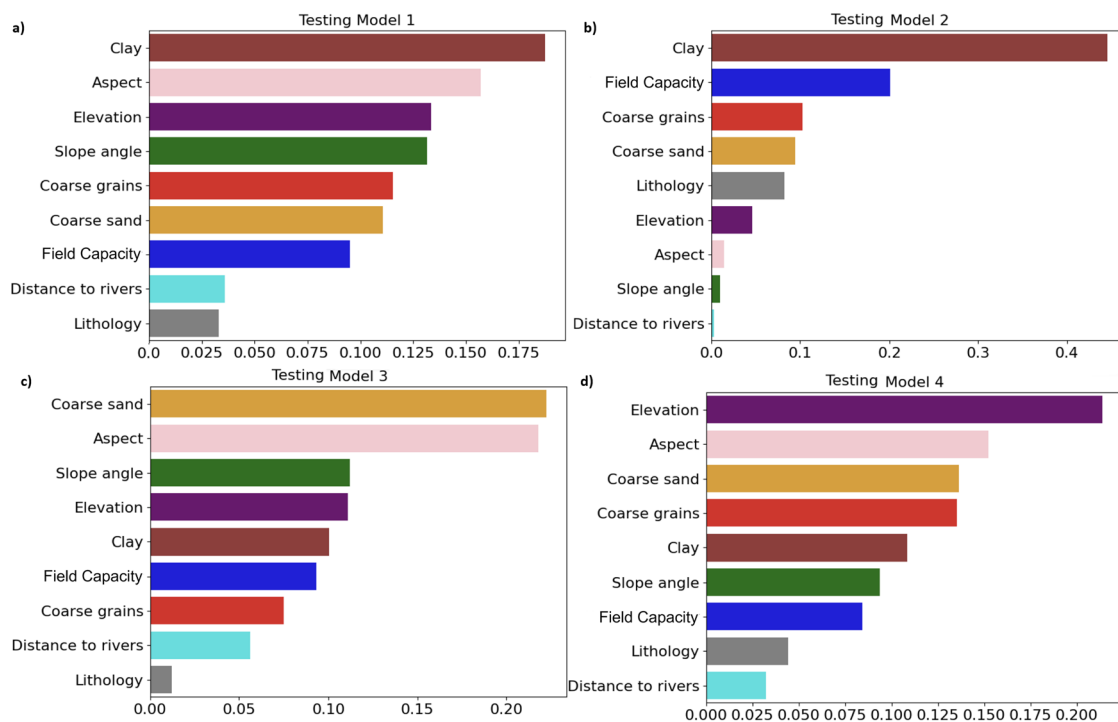


Fig. 8 Predisposing factors (y axis) importance scores (x axis) for the four testing models (a testing model 1; b testing model 2; c testing model 3; d testing model 4)

The third and fourth testing models used only points in the North region (testing model 3) and the Lisbon region (testing model 4) and were created with the aim of restricting the variation of the landslide predisposing factor values and identifying how such variation would affect the importance ranking. Testing model 3 ranked distance to rivers and lithology as the least important predisposing factors. Coarse sand was the most important predisposing factor in testing model 3 while clay content, which got first position in testing model 1, dropped to the fifth position. Testing model 4 also scored distance to rivers and lithology as the lowest important predisposing factors. Elevation, slope aspect, coarse sand, and coarse grains were the predisposing factors with higher score in testing model 4, showing similarities with the top-ranked predisposing factors of testing model 1.

Final phase

The final phase took place with the elaboration of the three final models considering three fixed critical rainfall durations. Figure 9 shows the feature importance of the three final models and enables the visualisation of how the influence of each feature changes according to the range of critical rainfall duration.

The final model 1 that used as dependent variable the critical daily rainfall intensities within 3 days of duration shows a higher importance on coarse grains and slope angle factors. The final model 2 trained with critical daily rainfall intensity considering the duration of 8 days gives the higher importance scores for slope aspect and coarse grains. Finally, final model 3 trained with critical daily rainfall intensities within 13 days of duration increases the clay content importance up to the top and gives the second

higher importance to field capacity, a feature that received little importance on the first two models. Lithology, distance to rivers, and coarse sand content in the soil received constantly low scores on all the three final models.

Aiming to increase the model explainability, it was also analysed the partial dependence of the three most important features for each of the final models. The partial dependence plots (Fig. 10) show how the dependent variable varies according to the change of one feature while the others are fixed, facilitating the interpretation of the RF results.

It is shown by the partial dependence plots that, for final model 1, the three most important features have a positive correlation to the dependent variable for the most part (slope angle starts to have a negative correlation after around 35 degrees). Coarse grains that have a strong positive correlation for the final model 1 shows a more complex behavior in relation to the dependent variable in final model 2. The clay content, identified as the most influential feature in final model 3, exhibits a negative correlation with the critical daily rainfall intensity values.

Model performance

The performance of the models was analysed with RMSE, MAE, and MAPE for the regression models, and the metrics accuracy and precision for the classifier model (Table 3).

The results disclosed that the testing models 1, 3, and 4 got a RMSE value around 3; MAPE values that range from 17 to 20% and MAE value around 2.5. Testing model 2 has a very high accuracy of 96%. The four testing models produced reasonable results and proved to

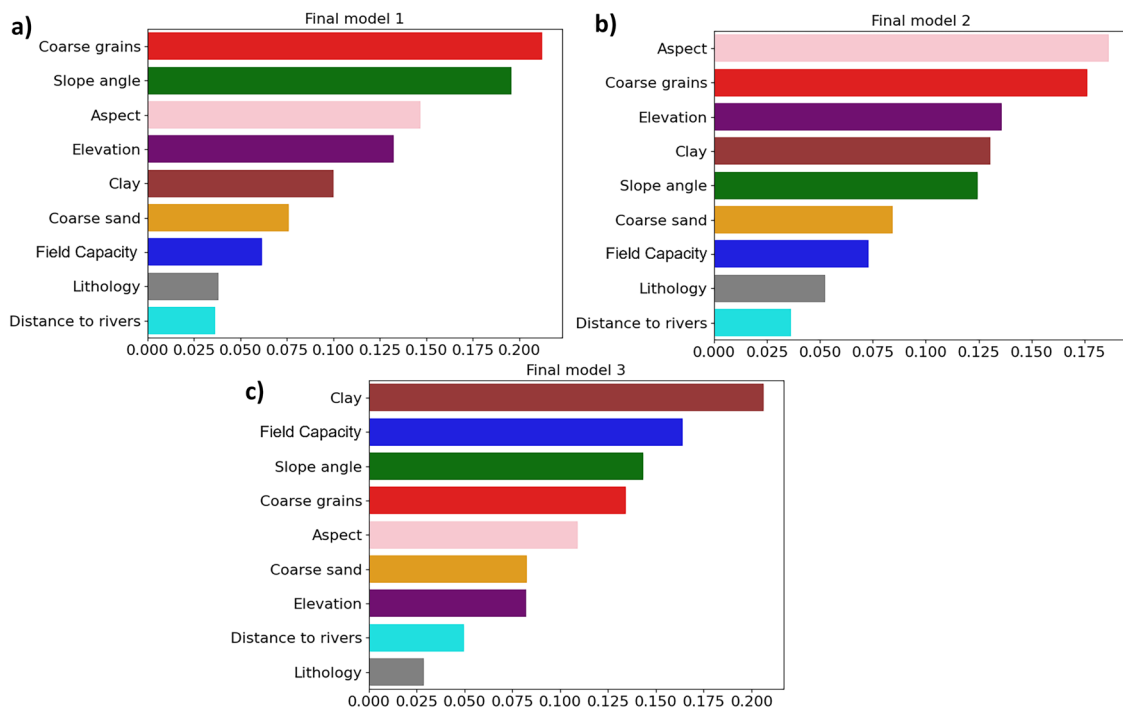


Fig. 9 Landslide predisposing factors (y axis) importance scores (x axis) for the final models (**a** final model 1; **b** final model 2; **c** final model 3)

be suitable for use in this work. The final models received similar results, with final model 2 achieving the highest level of performance.

Generation of the final maps

The final models were built employing the above-mentioned data (training datasets) together with testing and validating steps. The three final models were then used to assess the shallow landslide daily rainfall intensity threshold considering three critical rainfall durations (3, 8, and 13 days) at a resolution of 30 m/px (Fig. 11). Since all the study area covers a similar range of values of the landslide predisposing factors used for training the model, we considered that the model can be extrapolated to the complete mainland Portugal.

The maps shown in Fig. 11 represent a detailed distribution of daily rainfall threshold for shallow landslides for three different cases, and although some trends are evident, the heterogeneity of the values distribution is also clear. In general, for the maps generated with the final models 1 and 3, above the latitude 40 N, there is a higher accumulation of pixels with high values (green/blue colours), whereas southward this latitude most values are intermediate to low. However, there are areas of low rainfall intensity in the northern regions and of high rainfall intensity in the southern regions, contrary to the general trend. The map generated with the final model 2 shows a more homogeneous distribution with medium to higher values being predominant on almost all over the study area.

The heterogeneous distribution of daily rainfall intensity values across the map is the key to identifying the role of the predisposing factors in quantifying the rainfall intensity that triggers the shallow landslides. These results show the complexity of the terrain and

how a variety of factors can influence the occurrence of landslides. It should be highlighted that the maps of Fig. 11 do not analyse landslide susceptibility. Each pixel's rainfall intensity value acts as a critical threshold, whereby surpassing this threshold with observed rainfall is likely to trigger landslides.

Discussion

Influence of landslide predisposing factors

The exclusion of distance to fault in all four models suggests that, in our case study, shallow landslides may not be significantly influenced by geological structures. Instead, soil composition and terrain morphology appear to be the predominant factors. Alternatively, it could mean that shallow landslides, potentially affected by faults, do not exhibit distinguishable differences in critical rainfall intensities. Both hypotheses should be further confirmed with future analysis of shallow landslide susceptibility.

After running testing model 1, we found that the clay content had a high importance in the model as expected when comparing the clay content distribution (Fig. 5) and the critical daily rainfall intensity distribution (Fig. 7). Other 3 testing models were elaborated in order to assess if the method applied in this work is useful for giving insights of the correlation of landslide predisposing factors and critical daily rainfall intensity.

Testing model 2, which was built to differentiate the North group from the Lisbon group, ranked the predisposing factors related to morphology and topography with very low scores and used mainly the clay content to make the classifications. The fact that testing model 2 scored extremely low on the factors related to morphology, while the same factors are highly scored by testing models 1, 3, and 4, indicates that these factors may be only useful for assessing the

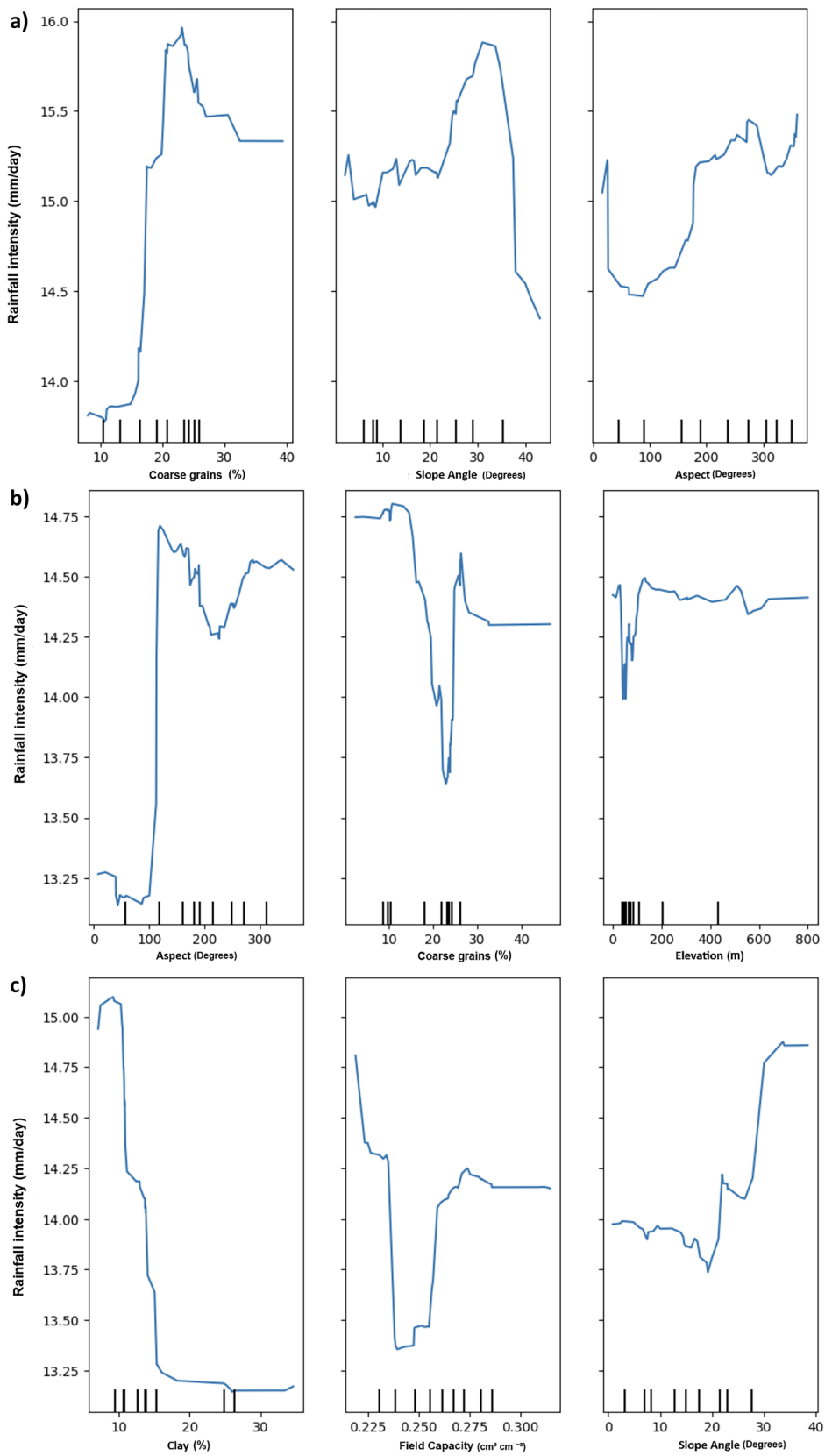


Fig. 10 Partial dependence for the three most influential landslide predisposing factors of the final models (**a** final model 1; **b** final model 2; **c** final model 3)

Table 3 RMSE, MAE, and MAPE values for regression models and accuracy/precision values for the classifier model

	Testing model 1	Testing model 2	Testing model 3	Testing model 4	Final model 1	Final model 2	Final model 3
Root mean squared error (RMSE)	2.9	-	3.3	2.6	2.9	3	2.6
Mean absolute error (MAE)	2.4	-	2.6	2.3	2.4	2.1	2.2
Mean absolute percentage error (MAPE)	17.2%	-	20.3%	17%	18%	15%	15.7%
Accuracy	-	0.96	-	-	-	-	-
Precision	-	0.95	-	-	-	-	-

critical daily rainfall intensity and not to differentiate geographical regions. Also, the fact that the clay content is relatively important for testing models 3 and 4, in which it analyses the two regions separately, indicates that the clay content is also being used to relate critical rainfall intensity with landslide predisposing factors, and not only to differentiate North and Lisbon groups.

The analysis and comparison of all four testing model results helped to clarify that the methodology used in this study has the potential to give insight into the relationship between predisposing factors and critical daily rainfall intensity.

In the final phase of the study, we analysed the overall and specific importance of predisposing factors for the three final models. The coarse grain content emerged as the most significant factor when considering a fixed rainfall duration of 3 days, but its relevance diminished as the duration increased. Interestingly, when we examined the partial dependence of coarse grain content on the final models 1 and 2, we found contrasting results. In the case of final model 1, the partial dependence of coarse grain content showed a positive correlation, whereas for final model 2, it exhibited a negative correlation. This discrepancy highlights the complexity of the landslide process. Additionally, field capacity received a high importance score only for final model 3, which utilised the longest critical rainfall duration.

Maybe during short-duration rainfall events, the soil may not have sufficient time to absorb water, regardless of its field capacity, which could explain the low importance of this feature in the final models 1 and 2.

The high importance received by soil properties (e.g. clay content and coarse grains) given by the final models could be explained by the fact that these predisposing factors may determine the infiltration capacity of the soil and thus will influence the amount of rain necessary to trigger shallow landslides.

Daily rainfall intensity threshold for shallow landslides

Figure 11 shows the final maps with the distribution of the daily rainfall intensity threshold considering three critical rainfall durations. These results can give insights into how the predisposing factors influence the spatial variation of the daily rainfall intensity threshold for shallow landslides in Portugal.

The distribution of different daily rainfall intensity thresholds across the entire study area confirms unconditionally that, if we consider critical rainfall event durations of 3 and 13 days, the northern region of Portugal needs higher rainfall intensity events to trigger shallow landslides, compared with regions under latitude 40 N, that are, in general, associated with lower rainfall intensities.

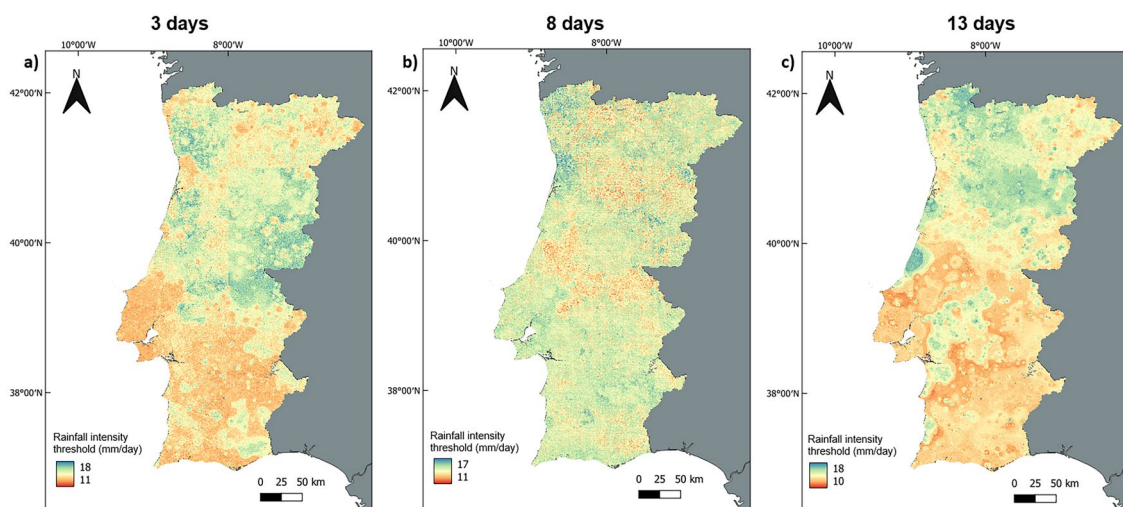


Fig. 11 Spatial distribution of the daily rainfall intensity thresholds considering three critical rainfall durations (a 3 days of critical rainfall duration; b 8 days of critical rainfall duration; c 13 days of critical rainfall duration)

These results show the different daily rainfall intensity threshold distributions follow a regional trend, and with the model predisposing factor ranking we know which predisposing factor could be influencing it. When considering critical rainfall durations of 8 days, the trend is not so evident at a national scale.

The areas with the highest rainfall intensity thresholds, when considering critical rainfall duration of 3 and 13 days, are generally concentrated in the region with the highest mean annual precipitation (Fig. 1b), meaning that the shallow landslides in the south part of the country can be triggered by lower rainfall intensities in absolute terms, in comparison to those registered in the northern region. That does not mean that regions related to a lower rainfall intensity threshold are necessarily more susceptible to landslides, since they are located, for the most part, in regions that receive a lower amount of precipitation per year.

Figure 11 also shows a heterogeneity in the distribution of daily rainfall intensity thresholds. Although there is a concentration of higher values in the north and lower values in the southern regions (for the final models 1 and 3), the models also predicted specific areas that differ from the trend. These heterogeneity distributions have been rarely shown in works about landslide thresholds.

The higher rainfall intensity thresholds concentrated in the area with the highest mean annual precipitation also tell us about landscape evolution and adaptation. In every geographical area, the geomorphology, especially the system of slope instability, tends to naturally attain an equilibrium state that aligns with the typical amount of rainfall experienced in that particular region (Pedrozzini 2004; Zêzere et al. 2015). Therefore, the slopes are adapted to the amount of precipitation it receives with time. It is important to highlight that, in order to fully understand the processes of shallow landslide triggering and to get insights about the susceptibility for the landslide occurrence in the area of study, it is necessary to analyse the map showing the rainfall intensity threshold together with the mean annual precipitation map.

Is it not possible to conclude with absolute certainty that the predisposing factors ranked by RF are the ones conditioning the spatial distribution of rainfall intensity thresholds (Stegar et al. 2021), but the different models analysed in this work provide some insights of possible correlations.

Since the distribution of mean annual precipitation in Portugal is heterogeneous, testing models 3 and 4 helped to better understand the role of predisposing factors in defining the rainfall intensity threshold, since they use groups of events that are located in areas receiving similar amounts of precipitation per year and within a limited variation of the predisposing factors (compared with testing model 1). Knowing that correlation does not mean causation (Stegar et al. 2021), this work does not exclude that the rainfall intensity thresholds could be influenced by other different factors, including anthropogenic, geological, climatological, and environmental aspects that have not been directly taken into account in the implemented models. Nevertheless, further studies may be useful to analyse the appropriate approaches for selecting additional predisposing factors and improve even more the models' performance and explainability.

The results of this study may have implications for landslide prediction and the generated model may be suitable for the effective reduction and mitigation of landslide risk in Portugal and for various regions worldwide characterised by similar geomorphological

settings. Many studies suggest that the definition of the warning threshold is very dependent on the selected rain gauge used to define the landslide rainfall threshold (Melillo et al. 2018; Abraham et al. 2020). Our method tries to overcome this issue by representing shallow landslide triggers as continuous values across the study area on a fine scale, rather than defining one rainfall threshold for a larger area, which could be beneficial for future warning systems. Further studies may be proposed to complement these findings using other novel machine learning and ensemble techniques in order to enhance our understanding of landslide triggers and improve landslide prediction.

It should be highlighted that the results are conditioned by the intrinsic uncertainty of the data collection, mainly regarding the filter used to classify the landslide occurrences as shallow or not. The developed filter does not provide certainty in the classification of the events; however, we hope that the method used reduces the chance of using noisy data in the models. Given the constraints of this study, particularly the limited dataset available for model development, we recommend several future validation measures to bolster our findings. These include acquiring an expanded dataset on shallow landslides, employing alternative machine learning or statistical methodologies, and applying our approach to different databases and geographical regions.

Conclusions

This study elaborated RF models to evaluate the influence of each landslide predisposing factor in assessing the daily rainfall intensity threshold for shallow landslides in Portugal. The RF models were employed with four different datasets, during the first initial phase, to assess the ability of the method to give insights about the correlation between predisposing factors and critical daily rainfall intensities. During the final phase, other three models were created to analyse the spatial variation of the daily rainfall intensity thresholds across mainland Portugal.

Moreover, the accuracy and validation of the models confirm that the derived maps for this study are effectively reliable. The main achievements of this work are as follows:

1. The landslide predisposing factors that were more useful to assess the daily rainfall intensity threshold in Portugal are the coarse grains, slope aspect, clay content, slope angle, and field capacity.
2. Above the latitude 40 N in Portugal, there is a higher concentration of pixels that represents the high rainfall intensity threshold for shallow landslides if considering critical rainfall durations of 3 and 13 days.
3. When considering critical rainfall durations of 3 and 13 days, the Southern part of Portugal is dominated by the low rainfall intensity threshold values, which is in line with the registered low mean annual precipitation.
4. When considering critical rainfall durations of 8 days, the trend of the spatial distribution of rainfall intensity threshold is not evident at the national scale.
5. Although earlier studies have addressed rainfall thresholds for landslide triggering in Portugal (Vaz 2021; Zêzere et al. 2015), our findings provide a more detailed exploration of spatial variations of rainfall thresholds.

The findings of this research contribute to the advancement of knowledge in landslide science and can provide valuable guidance for decision-makers and practitioners involved in landslide risk reduction efforts.

Author contribution

All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed with the contribution of all authors. The first draft of the manuscript was written by Caio Villaça and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

Open access funding provided by FCT|FCCN (b-on). This research was supported by Portuguese public funds through FCT—Fundação para a Ciência e a Tecnologia, I.P., by the Research Unit UIDB/00295/2020 (<https://doi.org/10.54499/UIDB/00295/2020>) and UIDP/00295/2020. Caio Villaça was funded through FCT, I.P., by the Grant 2022.14473.BD.

Data availability

The database regarding the occurrences can be found at: Zêzere, J. L., Pereira, S., Tavares, A. O., Bateira, C., Trigo, R. M., Quaresma, I., Santos, P. P., Santos, M., & Verde, J. (2022). DISASTER database on hydro-geomorphologic disasters in Portugal (Version 1, 2022) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.7117037>.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

Abraham MT, Satyam N, Rosi A, Pradhan B, Segoni S (2020) The selection of rain gauges and rainfall parameters in estimating intensity-duration thresholds for landslide occurrence: case study from Wayanad (India). *Water* 12(4):1000
 Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>

Brunetti MT, Peruccacci S, Rossi M, Luciani S, Valigi D, Guzzetti F (2010) Rainfall thresholds for the possible occurrence of landslides in Italy. *Nat Hazard* 10(3):447–458. <https://doi.org/10.5194/nhess-10-447-2010>
 Campbell RH (1975) Soil slips, debris flows, and rainstorms in the Santa Monica Mountains and Vicinity. Government Printing Office, Southern California. U.S
 Capitani M, Ribolini A, Bini M (2013) The slope aspect: a predisposing factor for landsliding? *CR Geosci* 345(11):427–438. <https://doi.org/10.1016/j.crte.2013.11.002>
 Catani F, Lagomarsino D, Segoni S, Tofani V (2013) Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. *Nat Hazard* 13(11):2815–2831. <https://doi.org/10.5194/nhess-13-2815-2013>
 Chen W, Peng J, Hong H, Shahabi H, Pradhan B, Liu J, Zhu A-X, Pei X, Duan Z (2018) Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County, Jiangxi Province, China. *Sci Total Environ* 626:1121–1135. <https://doi.org/10.1016/j.scitotenv.2018.01.124>
 Chen W, Xie X, Wang J, Pradhan B, Hong H, Bui DT, Ma J (2017) A comparative study of logistic model tree random forest and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena* 151:147–160
 Chen W, Sun Z, Han J (2019) Landslide susceptibility modeling using integrated ensemble weights of evidence with logistic regression and random forest models. *Applied Sciences* 9(1), Artigo 1. <https://doi.org/10.3390/app9010171>
 Cornes RC, van der Schrier G, van den Besselaar EJM, Jones PD (2018) An ensemble version of the E-OBS temperature and precipitation data sets. *Journal of Geophysical Research: Atmospheres* 123(17):9391–9409. <https://doi.org/10.1029/2017JD028200>
 Distefano P, Peres DJ, Scandura P, Cancelliere A (2022) Brief communication: Introducing rainfall thresholds for landslide triggering based on artificial neural networks. *Nat Hazard* 22(4):1151–1157
 Fang Z, Wang Y, van Westen C, Lombardo L (2024) Landslide hazard spatiotemporal prediction based on data-driven models: estimating where, when and how large landslide may be. *Int J Appl Earth Obs Geoinf* 126:103631
 Gómez H, Kavzoglu T (2005) Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin. *Venezuela Engineering Geology* 78(1):11–27. <https://doi.org/10.1016/j.enggeo.2004.10.004>
 Guzzetti F, Peruccacci S, Rossi M, Stark CP (2007) Rainfall thresholds for the initiation of landslides in central and southern Europe. *Meteorol Atmos Phys* 98(3):239–267. <https://doi.org/10.1007/s00703-007-0262-7>
 Lima P, Steger S, Glade T, Murillo-García FG (2022) Literature review and bibliometric analysis on data-driven assessment of landslide susceptibility. *J Mt Sci* 19(6):1670–1698
 Liu Z, Gilbert G, Cepeda JM, Lysdahl AOK, Piciullo L, Hefre H, Lacasse S (2021) Modelling of shallow landslides with machine learning algorithms. *Geosci Front* 12(1):385–393
 Melillo M, Brunetti MT, Peruccacci S, Gariano SL, Roccati A, Guzzetti F (2018) A tool for the automatic calculation of rainfall thresholds for landslide occurrence. *Environ Model Softw* 105:230–243
 Merghadi A, Yunus AP, Dou J, Whiteley J, ThaiPham B, Bui DT, Abderrahmane B (2020) Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth-Sci Rev* 207:103225
 Montrasio L, Valentino R (2008) A model for triggering mechanisms of shallow landslides. *Nat Hazard* 8(5):1149–1159. <https://doi.org/10.5194/nhess-8-1149-2008>
 Nocentini N, Rosi A, Segoni S, Fanti R (2023) Towards landslide space-time forecasting through machine learning: the influence of rainfall parameters and model setting. *Front Earth Sci* 11:1152130
 Nsengiyumva JB, Valentino R (2020) Predicting landslide susceptibility and risks using GIS-based machine learning simulations, case of upper Nyabarongo catchment. *Geomat Nat Haz Risk* 11(1):1250–1277. <https://doi.org/10.1080/19475705.2020.1785555>

- Palladino MR, Viero A, Turconi L, Brunetti MT, Peruccacci S, Melillo M, Luino F, Deganutti AM, Guzzetti F (2018) Rainfall thresholds for the activation of shallow landslides in the Italian Alps: the role of environmental conditioning factors. *Geomorphology* 303:53–67. <https://doi.org/10.1016/j.geomorph.2017.11.009>
- Pedrozi G (2004) Triggering of landslides in canton Ticino (Switzerland) and prediction by the rainfall intensity and duration method. *Bull Eng Geol Environ* 63:281–291
- Peruccacci S, Brunetti MT, Gariano SL, Melillo M, Rossi M, Guzzetti F (2017) Rainfall thresholds for possible landslide occurrence in Italy. *Geomorphology* 290:39–57. <https://doi.org/10.1016/j.geomorph.2017.03.031>
- Popescu ME (2002) Landslide causal factors and landslide remedial options. In 3rd international conference on landslides, slope stability and safety of infra-structures. CI-Premier PTE LTD Singapore, pp 61–81
- Pourghasemi HR, Kariminejad N, Amiri M, Edalat M, Zarafshar M, Blaschke T, Cerda A (2020) Assessing and mapping multi-hazard risk susceptibility using a machine learning technique. *Sci Rep* 10(1), Artigo 1. <https://doi.org/10.1038/s41598-020-60191-3>
- Ramos TB, Horta A, Gonçalves MC, Pires FP, Duffy D, Martins JC (2017) The INFOSOLO database as a first step towards the development of a soil information system in Portugal. *CATENA* 158:390–412. <https://doi.org/10.1016/j.catena.2017.07.020>
- Reichenbach P, Rossi M, Malamud BD, Mihir M, Guzzetti F (2018) A review of statistically-based landslide susceptibility models. *Earth Sci Rev* 180:60–91. <https://doi.org/10.1016/j.earscirev.2018.03.001>
- Rosi A, Frodella W, Nocentini N, Caleca F, Havenith HB, Strom A, Tofani V (2023) Comprehensive landslide susceptibility map of Central Asia. *Natural Hazards and Earth System Sciences* 23(6):2229–2250
- Steger S, Mair V, Kofler C, Pittore M, Zebisch M, Schneiderbauer S (2021) Correlation does not imply geomorphic causation in data-driven landslide susceptibility modelling—benefits of exploring landslide data collection effects. *Sci Total Environ* 776:145935
- Taalab K, Cheng T, Zhang Y (2018) Mapping landslide susceptibility and types using random forest. *Big Earth Data* 2(2):159–178
- Tehrani FS, Calvello M, Liu Z et al (2022) Machine learning and landslide studies: recent advances and applications. *Nat Hazards* 114:1197–1245. <https://doi.org/10.1007/s11069-022-05423-7>
- Valenzuela P (2017) Landslide triggering in Asturias (NW Spain): Rainfall and soil moisture conditions modelling. PhD Thesis, University of Oviedo, Oviedo
- Vaz T (2021) Movimentos de vertente desencadeados pela precipitação em Portugal continental. PhD Thesis, IGOT, University of Lisbon
- Vorpahl P, Elsenbeer H, Märker M, Schröder B (2012) How can statistical models help to determine driving factors of landslides? *Ecol Model* 239:27–39
- Wieczorek G, Guzzetti F (1999) A review of rainfall thresholds for triggering landslides
- Wu W, Sidle RC (1995) A distributed slope stability model for steep forested basins. *Water Resour Res* 31(8):2097–2110. <https://doi.org/10.1029/95WR01136>
- Zêzere JL, Trigo RM, Trigo IF (2005) Shallow and deep landslides induced by rainfall in the Lisbon region (Portugal): assessment of relationships with the North Atlantic Oscillation. *Nat Hazard* 5(3):331–344. <https://doi.org/10.5194/nhess-5-331-2005>
- Zêzere JL, Pereira S, Tavares AO, Bateira C, Trigo RM, Quaresma I, Santos PP, Santos M, Verde J (2014) DISASTER: A GIS database on hydro-geomorphologic disasters in Portugal. *Nat Hazards* 72(2):503–532. <https://doi.org/10.1007/s11069-013-1018-y>
- Zêzere JL, Vaz T, Pereira S, Oliveira SC, Marques R, Garcia RAC (2015) Rainfall thresholds for landslide activity in Portugal: a state of the art. *Environmental Earth Sciences* 73(6):2917–2936. <https://doi.org/10.1007/s12665-014-3672-0>
- Zêzere JL, Pereira S, Melo R, Oliveira SC, Garcia RAC (2017) Mapping landslide susceptibility using data-driven methods. *Sci Total Environ* 589:250–267. <https://doi.org/10.1016/j.scitotenv.2017.02.188>

Caio Villaça (✉) · **Pedro Pinto Santos** · **José Luís Zêzere**

Centre of Geographical Studies, TERRA Associate Laboratory. IGOT, University of Lisbon, Lisbon, Portugal
Email: caiovnv10@gmail.com