Earth-Science Reviews 258 (2024) 104927

Contents lists available at ScienceDirect

Earth-Science Reviews

journal homepage: www.elsevier.com/locate/earscirev

A benchmark dataset and workflow for landslide susceptibility zonation

Massimiliano Alvioli [a,*], Marco Loche [b,c], Liesbet Jacobs [d], Carlos H. Grohmann [e], Minu Treesa Abraham [f,g], Kunal Gupta [h], Neelima Satyam [h], Gianvito Scaringi [b], Txomin Bornaetxea [a,i], Mauro Rossi [a], Ivan Marchesini [a], Luigi Lombardo [j], Mateo Moreno [j,k], Stefan Steger [l], Corrado A.S. Camera [m], Greta Bajni [m], Guruh Samodra [n], Erwin Eko Wahyudi [o], Nanang Susyanto [p], Marko Sinčić [q], Sanja Bernat Gazibara [q], Flavius Sirbu [r], Jewgenij Torizin [s], Nick Schüßler [s], Benjamin B. Mirus [t], Jacob B. Woodard [t], Héctor Aguilera [u], Jhonatan Rivera-Rivera [u,v]

[a] *Consiglio Nazionale delle Ricerche, Istituto di Ricerca per la Protezione Idrogeologica, via Madonna Alta 126, I-06128 Perugia, Italy*
[b] *Institute of Hydrogeology, Engineering Geology and Applied Geophysics, Charles University, Albertov 6, 128 43 Prague, Czech Republic*
[c] *Institute of Rock Structure and Mechanics, Czech Academy of Sciences, V. Holešovičkách 41, 182 09 Prague, Czech Republic*
[d] *Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, the Netherlands*
[e] *Institute of Astronomy, Geophysics and Atmospheric Sciences, Universidade de São Paulo, Rua do Matão 1226, 05508-090, São Paulo, SP, Brazil*
[f] *Methods for Model–based Development in Computational Engineering, RWTH Aachen, 52062, Germany*
[g] *Norwegian Geotechnical Institute, 0484 Oslo, Norway*
[h] *Department of Civil Engineering, Indian Institute of Technology, Indore, India*
[i] *Departamento de Geografía, Prehistoria y Arqueología, Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU), Vitoria-Gasteiz 01006, Spain*
[j] *Faculty of Geo–Information Science and Earth Observation (ITC), University of Twente, PO Box 217, Enschede AE 7500, the Netherlands*
[k] *Center for Climate Change and Transformation, Eurac Research, Bolzano, Italy*
[l] *GeoSphere Austria, Vienna, Austria*
[m] *Dipartimento di Scienze della Terra "A. Desio", Università degli Studi di Milano, Milan, Italy*
[n] *Department of Environmental Geography, Faculty of Geography, Universitas Gadjah Mada, Indonesia*
[o] *Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Indonesia*
[p] *Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Indonesia*
[q] *Department of Geology and Geological Engineering, Faculty of Mining, Geology and Petroleum Engineering, University of Zagreb, Pierottijeva 6, HR-10000 Zagreb, Croatia*
[r] *Institute for Advance Environmental Research, West University of Timisoara, Oituz 4, 300086 Timişoara, Romania*
[s] *Federal Institute for Geosciences and Natural Resources, Hannover, Germany*
[t] *U.S. Geological Survey, Geologic Hazards Science Center, Golden, CO, United States of America*
[u] *Geological Survey of Spain (IGME-CSIC), Rios Rosas 23, 28003 Madrid, Spain*
[v] *ETSI Topografía. Geodesia y Cartografía, Universidad Politécnica de Madrid (UPM), 28031 Madrid, Spain*

ARTICLE INFO

Keywords:
Landslide susceptibility
Benchmark dataset
Machine learning
Statistical modeling
Landslide susceptibility mapping
Slope units
Landslide inventory
Geomorphological mapping
Spatial analysis
Geomorphometry

ABSTRACT

Landslide susceptibility shows the spatial likelihood of landslide occurrence in a specific geographical area and is a relevant tool for mitigating the impact of landslides worldwide. As such, it is the subject of countless scientific studies. Many methods exist for generating a susceptibility map, mostly falling under the definition of statistical or machine learning. These models try to solve a classification problem: given a collection of spatial variables, and their combination associated with landslide presence or absence, a model should be trained, tested to reproduce the target outcome, and eventually applied to unseen data.

Contrary to many fields of science that use machine learning for specific tasks, no reference data exist to assess the performance of a given method for landslide susceptibility. Here, we propose a benchmark dataset consisting of 7360 slope units encompassing an area of about $4,100$ km^2 in Central Italy. Using the dataset, we tried to answer two open questions in landslide research: (1) what effect does the human variability have in creating susceptibility models; (2) how can we develop a reproducible workflow for allowing meaningful model comparisons within the landslide susceptibility research community.

* Corresponding author.
E-mail address: massimiliano.alvioli@irpi.cnr.it (M. Alvioli).

https://doi.org/10.1016/j.earscirev.2024.104927
Received 20 February 2024; Received in revised form 30 August 2024; Accepted 7 September 2024
Available online 11 September 2024

With these questions in mind, we released a preliminary version of the dataset, along with a "call for collaboration," aimed at collecting different calculations using the proposed data, and leaving the freedom of implementation to the respondents. Contributions were different in many respects, including classification methods, use of predictors, implementation of training/validation, and performance assessment. That feedback suggested refining the initial dataset, and constraining the implementation workflow. This resulted in a final benchmark dataset and landslide susceptibility maps obtained with many classification methods.

Values of area under the receiver operating characteristic curve obtained with the final benchmark dataset were rather similar, as an effect of constraints on training, cross–validation, and use of data. Brier score results show larger variability, instead, ascribed to different model predictive abilities. Correlation plots show similarities between results of different methods applied by the same group, ascribed to a residual implementation dependence.

We stress that the experiment did not intend to select the "best" method but only to establish a first benchmark dataset and workflow, that may be useful as a standard reference for calculations by other scholars. The experiment, to our knowledge, is the first of its kind for landslide susceptibility modeling. The data and workflow presented here comparatively assess the performance of independent methods for landslide susceptibility and we suggest the benchmark approach as a best practice for quantitative research in geosciences.

1. Introduction

Landslide susceptibility assessment with statistical and machine learning methods requires a substantial amount of topographic, geomorphological, and environmental data to train and test a specific model. The combination of input data and the method of choice are the ingredients to prepare a classification of the study area based on specific mapping units - *i.e.*, elementary portions of the area. A simple square grid may be effective to discretize spatially distributed variables, but slope units work substantially better in representing topographic units with uniform likelihood of landslides to occur. Use of slope units, which are portions of the domain separated by drainage and divide lines, has conceptual advantages (Carrara et al., 1991; Alvioli et al., 2016; Jacobs et al., 2020; Rolain et al., 2023), although their delineation represented a practical limitation for a long time (Reichenbach et al., 2018).

Relevant input data, usually referred to as "predictors," "factors," "covariates," "features," or "independent/explanatory variables," are a mixed set of morphometric quantities and a variety of thematic data. A landslide inventory is also needed, representing the "dependent," "target," or "response" variable to be reproduced by the model (Guzzetti et al., 2012). Different landslide inventories may lead to different susceptibility maps (Bordoni et al., 2020; Pokharel et al., 2021; Bajni et al., 2022; Bornaetxea et al., 2023a; Dias and Grohmann, 2024), implying that a given inventory must be selected for a benchmark, which is a common dataset used to compare models developed by different users.

The choice of a specific method/model used to obtain a susceptibility map depends on software availability, personal background, and existence of relevant literature for the area of interest. New methods are proposed regularly; however, due to the lack of a benchmark dataset, it is difficult to judge the relative performance of methods across different studies, and the methods' applicability in a specific area (Süzen and Doyuran, 2004; Yesilnacar and Topal, 2005; Akgun, 2012; Thai Pham et al., 2020; Tien Bui et al., 2016; Merghadi et al., 2020).

Many publications exist on the subject of landslide susceptibility (Reichenbach et al., 2018; Merghadi et al., 2020), and several updated reviews about landslide susceptibility methods appear every year (*e.g.*, Lee, 2019; Shano et al., 2020; Dias et al., 2021; Das et al., 2022; Yong et al., 2022; Liu et al., 2022, 2023). For example, Lima et al. (2022) presented a bibliographic–oriented summary of the landslide susceptibility literature, Huang et al. (2024) centered their review on data–construction strategies, and Budimir et al. (2015) focused on aspects related to the choice of the common predictor sets. However, a meaningful comparison of many different methods requires a common benchmark dataset to train and test each of them in a systematic way. This is a standard procedure in machine learning science and practice: benchmark datasets exist for medical sciences, image recognition, linguistics and, in general, any data–driven research. The "Iris dataset" is a famous example of a benchmark in classification of numerical data into three different variants of the flower Iris (Fisher, 1936). Examples of open datasets for this purpose, listed in alphabetical order, are available on GitHub (2022); another example is the University of California Irvine machine learning repository (UCI, 2024).

Despite machine learning having been applied to landslide studies, no benchmark dataset exists for machine learning applications in landslide susceptibility. Benchmark data exist in other fields of the Geosciences. For example, Buiter et al. (2016) and Schreurs et al. (2016) benchmarked numerical models of thrust wedges with brittle materials, Kirby et al. (2022) discussed comparison of models for landslide tsunami generation, and Leung et al. (2024) described a benchmarking exercise for rock mass discontinuity mapping.

Here, we propose a benchmark dataset and describe the results of an array of different methods, using a consistent workflow, which can be a standard reference for landslide susceptibility studies. Of note, the benchmark dataset and associated comparison of methods were developed with input from many experts active in the field of landslide science.

This study was carried out over the course of about one year, and we summarize it as follows: (1) introducing a preliminary dataset to compare the outcome of different methods for landslide susceptibility assessment in a meaningful way, (2) collecting contributions of researchers active in the field of landslide susceptibility to use such dataset with their method of choice, (3) reviewing results and feedback of contributors, (4) revising the dataset to obtain a final benchmark dataset, distributed again with prescriptions about the workflow to obtain landslide susceptibility maps (LSMs), and (5) collecting final calculation results from different contributors and making them publicly available.

This work is the first attempt to achieve model comparability beyond the result provided by individual research groups but rather as a community effort. For this reason, our contribution aims at addressing a different task as compared to existing reviews. We certainly include elements of literature review, but these are systematically presented through a practical experiment run by as many research groups as we could gather. This process has undergone a two–step procedure to first explore differences among different methods, and then bring together the participating groups and devise a common modeling protocol. This procedure has ultimately been translated not only into a shared list of modeling recommendations but also into a shared dataset.

2. Review of essential literature on landslide susceptibility

Studies on landslide susceptibility have benefited from several notable technological advances over the last five decades. The earliest digital document on landslide susceptibility dates back to the 1970s, where Brabb et al. (1972) provided an expert–based susceptibility classification of a study site in California, United States, on the basis of his expectation whether a slope could be prone or not to fail. Since then,

the literature has witnessed a radical change in the spectrum of possible answers to the same question. Specifically, the following decades welcomed contributions that explored more numerical–oriented approaches.

With the advent of geographical information system software, information on landscape characteristics associated with potential failures became available in digital form. This first era of innovation produced a wave of contributions centered around heuristic models (Luckman, 1987; Leoni et al., 2009). These were later replaced by bivariate statistical approaches. Among these, tools such as certainty factor (Wislocki and Bentley, 1991) or weight of evidence (WoE – Bonham-Carter et al., 1990; Agterberg et al., 1990; Luzi et al., 2000) offered good performance and easy to implement models that were used until recent times (*e.g.*, Regmi et al., 2010). However, these methods suffered from a lack of quantitative outputs, which diminished model interpretability.

Statistical and machine learning susceptibility models provided a more quantitatively robust method of estimating landslide susceptibility, which led to its dominance in the field up to the present. The introduction of binomial statistical model by Atkinson and Massari (1998) increased susceptibility model performance and interpretation (Huang et al., 2020). Nevertheless, these models only allow for the relations between dependent and independent variables to be estimated in a linear fashion, an assumption that may not hold in many cases. For this reason, nonlinear extensions such as generalized additive models (GAMs – Brenning, 2008) have superseded them.

In the early 2000s, machine learning also made its appearance in the landslide susceptibility science, with a number of applications that extended widely to encompass (i) tree–based models (Yeon et al., 2010) and their three main derivatives: random forest (RF – Ho, 1995; Breiman, 2011; Hong et al., 2019), boosted regression trees and XGBoost (Zeng et al., 2023), ii) support vector machines (Huang and Zhao, 2018), (iii) artificial neural networks (ANNs – Amato et al., 2023; Bragagnolo et al., 2020). Historically, statistical and machine learning approaches occupied very distinct areas, with the former being sought after for its interpretability and uncertainty estimation components (Di Napoli et al., 2023), whereas the latter was used in performance–oriented applications (Marjanović et al., 2011). Only recently have these differences became more blurred, with statistical models incorporating spatial dependence information (Chalkias et al., 2020) and machine learning approaches offering tools to facilitate their interpretation (Dahal and Lombardo, 2023).

Notably, with the introduction of all these models, the landslide susceptibility community entered a somewhat dormant era between the years 2010 and 2020. During this time, a plethora of publications aimed at comparing certain models against others, each time taking a different set of tools under consideration and a different dataset to build such comparison. As a result, hundreds of articles appeared with no explicit research question other than comparing a set of models and a set of data in a particular context or setting. This practice does not allow for the systematic comparability required for general advancements in the landslide susceptibility field, and it is precisely with this idea in mind that the present work proposes a standard for a benchmark dataset (refer to Section 3).

In addition to exploring the effectiveness of different susceptibility model types, much work has investigated the effects of the ratio of landslide presence and absence data. The standard definition of landslide susceptibility, as given by Carrara et al. (1995) or Guzzetti et al. (1999), does not formally require retrieving an absolute probability as it is often the case in statistics. Conversely, the susceptibility definition corresponds to a relative probability between different mapping units. In other words, susceptibility assessment seeks to define which locations are more prone than others to experience a slope failure rather than assigning them an exact probability value (Akgun et al., 2008; Sterlacchini et al., 2011).

To strictly compute probabilities, a model should be fit with all the available presence/absence data. However, the common approach in the literature is to keep all the presences while subsampling the absences (Huang et al., 2024). Many examples can be found where a balanced sampling strategy is pursued (*e.g.*, Erener et al., 2017; Lucchese et al., 2021). In other contributions, the absences are still subsampled from the whole dataset but kept at a greater proportion with respect to the presences (Heckmann et al., 2014; Moreno et al., 2024; Bornaetxea et al., 2018). Importantly, sampling strategies vary between machine learning and statistical modeling.

If we consider statistical modeling, the implication of varying the proportion of the presence/absence label can essentially be seen in the global intercept. The first work to refer to this effect is by Petschko et al. (2014). There, the authors note the effect of sampling a subset of the absences on the global intercept and propose an equation to correct for this effect, thus effectively bringing the obtained relative probabilities to the standard strictly prescribed in statistics. The same line of research has been further explored (refer to the supplementary materials in Lombardo and Mai, 2018), where the effects on the global intercept have been demonstrated in a simulation exercise. In short, a dataset with much fewer presences than absences would estimate very negative global intercept values. This, in turn, applies a constant probability shift towards the left side of the susceptibility distribution. In other words, a balanced sampling choice returns probability values shaped according to a Gaussian or near–Gaussian distribution centered at around 0.5. Thus, as the absence proportion progressively increases, the distribution of susceptibility values becomes more and more positively skewed (Lombardo and Tanyas, 2022). A more positively skewed or even heavy–tailed susceptibility distribution matches the reality, with few locations being highly susceptible and most of the landscape is considered stable (Jia et al., 2021). The artificial transposition of this shape towards a normal distribution implies that any landscape is approximately split into two sides, 50 % to be considered stable and 50 % to be considered unstable. This is obviously not what happens in reality and it is also the reason why susceptibility values are almost always presented in a reclassified form. In such a way, grouping probability into low–to–high susceptibility classes removes the differences induced by values concentrated either in the bulk or tails of the distribution. This is, therefore, another area where many differences exist in the literature. In this sense, the Jenks natural break classification is quite common (Mărgărint et al., 2013; Elia et al., 2023), and alternatives can be found in an equal interval (Kavzoglu et al., 2014; Chen et al., 2016) or quantile descriptions of the susceptibility range (Steger et al., 2020; Wang et al., 2022).

The sampling strategies using machine/deep learning tools are more regulated. Machine learning largely prescribes that users select balanced sampling strategies (*e.g.*, Batista et al., 2004), unless custom–made loss functions are used to account for data imbalance (Prakash et al., 2020; Dahal et al., 2024). For this reason, many fewer studies explore absence selection effects (*e.g.*, Hong et al., 2019; Liang et al., 2021; Rabby et al., 2023).

3. Methods and data

The first action of this study was devising a tentative dataset, and publishing a call for expressions of interest in participating in a quantitative experiment comparing susceptibility methods on a proposed benchmark dataset. We proposed this experiment as a topical session at the annual European Geosciences Union General Assembly 2023.[1] Participants presented their calculations to obtain landslide susceptibility in the study area using the proposed dataset. The approaches of the 11 participating groups were different in many respects. In Sections 4.1–4.11, we report for each participant group information about (i) type of model, (ii) variable selection, (iii) calibration/validation approach, and (iv) performance assessment. Section 4.12 summarizes

[1] https://meetingorganizer.copernicus.org/EGU23/session/47046.

similarities and differences of the participants' results.

The second set of actions, collectively aimed at devising a final dataset, was based on the results of the previous step and is described in Section 5. Feedback from the previous step suggested that the dataset should be updated to remove collinearity, which was an issue for most of the contributions. Moreover, it was clear that the workflow applied to obtain LSMs should be standardized both for a meaningful comparison of results from different methods and for benchmarking independent calculations against the results presented here. To this end, a final benchmark dataset was obtained adding new predictors and removing collinearity by reducing the number of variables, as described in Section 5.1. The updated data were distributed to the contributors, with well–defined requirements for cross–validation (CV), so that the individual groups produced a more informative set of results.

3.1. Data

We selected a slope–unit (SU)–based dataset because SUs have a meaningful correspondence with topography (Guzzetti et al., 2006). We extracted a subset of the dataset used by Loche et al. (2022) for landslide susceptibility maps in Italy, who adopted an SU set previously optimized for Italy by Alvioli et al. (2020).

Out of the entire SU map of Italy, containing about 330,000 polygons, we selected a subset of 7360 units encompassing an area of $4,095\ \text{km}^2$ in Central Italy. Fig. 1 shows the spatial location of the area of interest. The data had an attribute table containing several different morphometric and thematic variables. The morphometric variables were calculated using the European digital elevation model EU–DEM with 25–m resolution.[2] A few variables were obtained from the SoilGrids global dataset (Hengl et al., 2017). Table 1 lists the full set of variables.

The SoilGrids dataset is an application of machine–learning models trained on over 230,000 soil profile observations from the world soil information WoSIS database (ISRIC, 2024)). Lower and upper limits of a 90 % prediction interval quantify prediction uncertainty. Global datasets and models are increasingly being used to make use of data–hungry, high–performance, large–scale, machine–learning models. The accuracy of global datasets depends on the density and quality of data points used for building the models. Freely available global products aer useful both in the context of this work, aiming at becoming a reference dataset for landslide susceptibility mapping, and for similar datasets, developed in different areas.

The original landslide location map in Loche et al. (2022) contained eight different presence/absence flags, corresponding to the point locations (highest point of landslide crown) of eight types of landslides from the Italian National landslide database assembled by the Italian Geological Survey (ISPRA; Trigila et al. (2010)). For this work, we selected only presence/absence of translational landslides.

To provide the contributors with two different landslide presence scenarios, we flagged landslide presence with two attribute fields, called `p1` and `p2`, which is similar to flagging an SU as unstable if it contains a minimum landslide area (Guzzetti et al., 2006; Schlögel et al., 2018).

To define `p1`, we selected SUs labeled as "without landslide" (`p1` flag: 0) where an SU contained no points at all, in 3766 cases ($1,443.1\ \text{km}^2$), and as "with landslides" (`p1` flag: 1) in the remaining 3594 cases ($2,652.1\ \text{km}^2$). For `p2`, we selected SUs labeled as "without landslides" (`p2` flag: 0) where an SU contained up to one point, in 5089 cases ($2,087.1\ \text{km}^2$), and as "with landslides" (`p2` flag: 1) in the remaining 2271 cases ($2,008.2\ \text{km}^2$).

Note that, using `p1` as landslide presence, one would have an approximately balanced dataset with respect to the number of zeros/ones; using `p2`, instead, one would have an approximately balanced dataset with respect to the total surface area covered by the SUs labeled either with zero, or one. Fig. 2 shows the spatial distribution of SUs labeled as positive/negative in the two cases. In such a varied methodological landscape pertaining to sampling ratios (Section 2), a detailed exploration of the selection of non–landslide data is beyond the scope of this work.

We invited participants to consider both landslide presence flags to produce two different LSMs for the study area. Moreover, we invited them to use their best strategy, or the strategy that best fits their model of choice, to produce a result for a landslide susceptibility index – a float number ranging from zero to unity – and an associated uncertainty, where possible.

4. Preliminary assessment of the benchmark dataset for landslide susceptibility

The following describe the methods applied in step one of the experiment, by each participating group. In each contribution, we have distinguished model selection, variable selection, calibration–validation approach, and model evaluation. In cases where no exclusion of variables is described, all variables were retained for the group's results.

4.1. Group 1. Application of the LAND–SUITE multi–model software

This contribution was presented as the (EGU) abstract by Bornaetxea et al. (2023b).

Model selection. Group 1 (G1) utilized the LAND–SUITE software (Rossi et al., 2022), a suite of R script modules (R Core Team, 2021) designed to support the landslide susceptibility inference process. LAND–SUITE contains several statistically driven approaches, including linear discriminant analysis (LDA), logistic regression (LR), and quadratic discriminant analysis (QDA). Additionally, the software provides an option to combine the outputs of the selected statistical methods into a single combination forecast model (CFM), where LR is used to determine the best fit among the original outputs (Rossi and Reichenbach, 2016). They tested all of the mentioned approaches (LDA, LR, QDA, and CFM), considering the two proposed landslide presence scenarios (`p1` and `p2`), along with the explanatory variables, resulting in eight LSMs.

Variable selection. Each modeling process was preceded by an exploration phase to identify possible correlations among pairs of explanatory variables and, in such cases, to select only the most significant variable. Including highly correlated variables in the model training would likely inflate the model error and uncertainty estimate, thus negatively affecting the overall performance and the interpretation of variable effects (Amato et al., 2019). This in turn would increase computational time and, in some extreme collinearity cases, may even hinder the model convergence, especially for statistical models. To address this, G1 computed mutual correlation coefficients among the 26 explanatory variables (including SU area) and considered two variables as highly correlated if the Pearson correlation coefficient exceeded $|0.7|$. A leave–one–out (LOO) test assessed the individual significance of each variable with respect to the others (Gong, 2006; Sin Yin et al., 2010). With n preliminary runs of each model, excluding one variable at a time, G1 determined which variable's absence resulted in the largest loss in model performance, and they excluded the less significant variable from each highly correlated pair. This approach is referred to in different ways depending on the user background, including terms such as jackknife tests in ecology (Shcheglovitova and Anderson, 2013), or ablation studies in computer science (Aguilera et al., 2022). Additionally, for the LR outputs, G1 verified the p–value corresponding to each variable. Variables with p–values $\gg 0.05$ were excluded from the analysis. This process was performed for each of the provided target variables (`p1` and `p2`). The original values of the explanatory variables were scaled between their minimum/maximum values.

Calibration–Validation approach. Every experiment (LDA, QDA, LR, and CFM) used the same training and validation data partition, which

[2] https://www.eea.europa.eu.

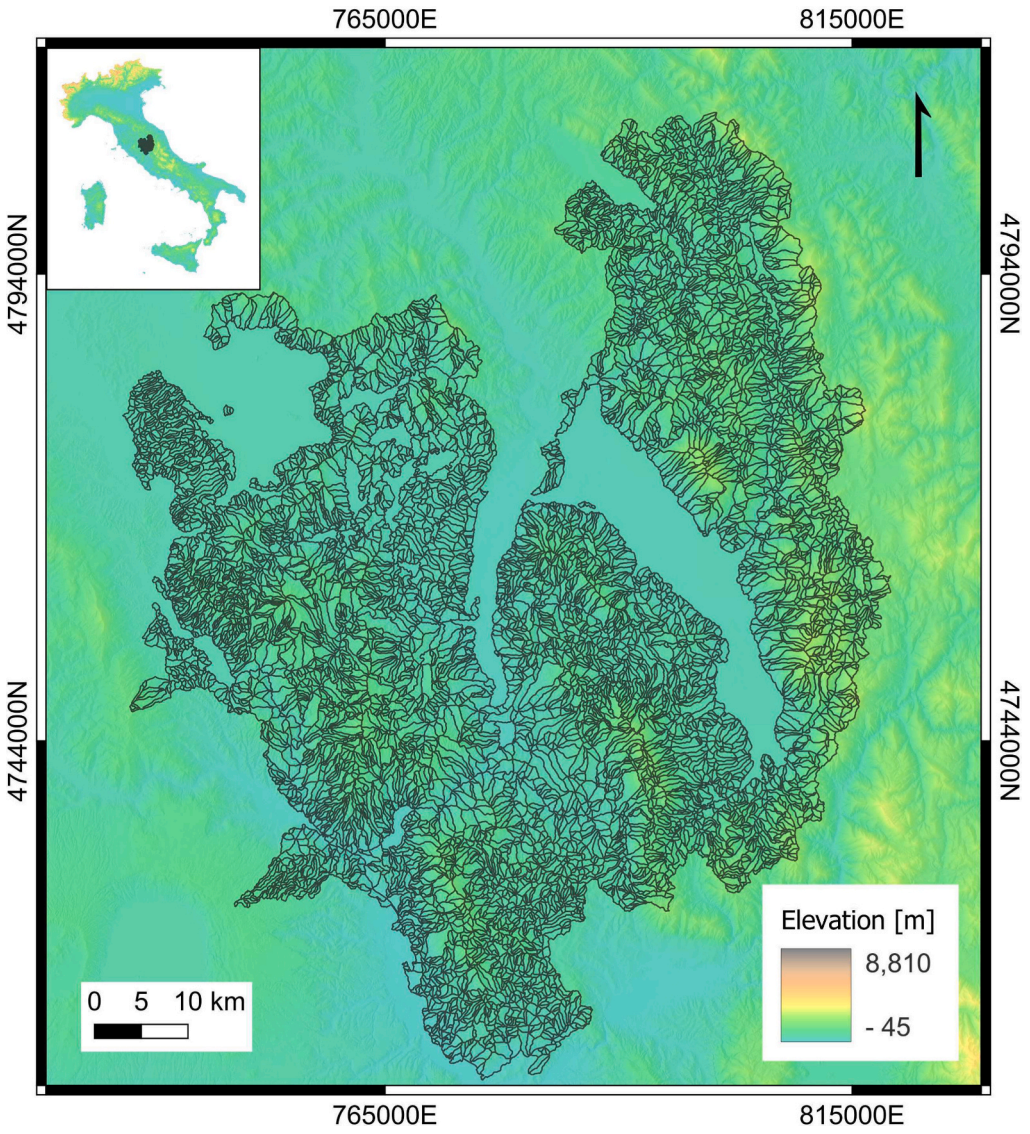


Fig. 1. Geographical location (inset) of the area covered by the slope unit set (main figure) selected in this work as a benchmark dataset for landslide susceptibility zonation. The dataset is a subset of the slope unit map obtained by Alvioli et al. (2020), and used by Loche et al. (2022) for a landslide susceptibility map of Italy. In the dataset proposed here, we selected point locations of translational landslides from the Italian national inventory known as 'IFFI' (Trigila et al., 2010). Map is in EPSG:32632 projected reference system.

involved a simple (one–fold) random CV approach. Group 1 allocated 75 % of the dataset for training and the remaining 25 % for validation. They made sure that a balanced number of landslide presence (1) and absence (0) were found in both the training and validation sets. To assess the internal variability of the results due to the randomly obtained training samples, G1 used the bootstrap resampling method (Davison and Hinkley, 1997). They conducted 100 resample iterations for each tested model and plotted the average and standard deviation of the results on variability plots (Rossi and Reichenbach, 2016).

Model evaluation. Validation of the results was performed using the area under the receiver operating characteristic (ROC) curves (AUC_{ROC}, Fawcett (2006)), calculated for both the training and validation samples. Group 1 obtained four–fold and histogram plots to visualize the overall agreement of the model compared to the observed results in the validation dataset. In all figures and tables, results corresponding to this paragraph are labeled as LDA, QDA, LR_1, and CFM (*cf.* Fig. 3).

4.2. Group 2. Generalized additive models with shrinkage option and geomorphological plausibility check

This contribution was presented as the EGU abstract by Camera and Bajni (2023).

Model selection. Group 2 (G2) applied GAMs, using the `mgcv` library in R (Wood, 2017). This class of models was selected because these models are easily interpretable and widely applied in recent literature with good results (*e.g.*, Goetz et al. (2011); Bajni et al. (2023); Fang et al. (2024); Wang et al. (2024)).

Variable selection. An exploratory correlation analysis was carried out between the 27 independent variables (SU area included). Variable selection was done during the GAM fitting through shrinkage, which consists in removing the variables that explain a small part of model variance (usually variables highly correlated with others). This approach is quite intuitive in the linear case, with a penalization term used to shrink the regression coefficient values towards zero, checking at each time whether the shrinkage leads to loss in performance or not (Ranstam and Cook, 2018). In a nonlinear case, the penalization is executed in two dimensions, both for the regression coefficients as well

Table 1
Variables in the attribute table of the preliminary dataset. In the table, SD stands for standard deviation and SU stands for slope unit. Depth to bedrock, bulk density, percentage weight of clay, sand and silt particles are from Hengl et al. (2017).

Column name	Variable	Short name
id	Unique slope unit identifier	–
slope_aver	Mean Slope Steepness [deg]	Sav
slope_stdd	SD of Slope within SU [deg]	Ssd
pcurv_aver	Mean Planar Curvature	PCav
pcurv_stdd	SD of Planar Curvature	PCsd
tcurv_aver	Mean Profile Curvature	TCav
tcurv_stdd	SD of Profile Curvature	TCsd
nthns_aver	Mean Northerness	NTav
nthns_stdd	SD of Northerness	NTsd
easns_aver	Mean Easterness	EAav
easns_stdd	SD of Easterness	EAsd
elev_avera	Mean Elevation [m]	ELav
elev_stddd	SD of Elevation [m]	ELsd
twi_averag	Mean Topographic Wetness Index	TWav
twi_stddev	SD of Topographic Wetness Index	TWsd
BDRICM_ave	Mean Depth to bedrock (<2.4 m) [cm]	BDRav
BDRICM_std	SD of Depth to bedrock [cm]	BDRsd
BLDFIE_ave	Mean Bulk density [kg/m^3]	DLBav
BLDFIE_std	SD of Bulk density [kg/m^3]	DLBsd
CLYPPT_ave	Mean Weight % of clay particles	CLYav
CLYPPT_std	SD of Weight % of clay particles	CLYsd
SNDPPT_ave	Mean Weight % of sand particles	SLTav
SNDPPT_std	SD of Weight % of sand particles	SLTsd
SLTPPT_ave	Mean Weight % of silt particles	SNDav
SLTPPT_std	SD of Weight % of silt particles	SNDsd
Max_Distan	Maximum Distance within SU [m]	MaxD
D_sqrt_A	Maximum Distance/$\sqrt{SU\,Area}$	DsqrA
presence1	Binary landslide presence flag	p1
presence2	Binary landslide presence flag	p2
area	Area [km^2]	–

as across the spline applied to the domain of the explanatory variable of interest (Wood and Augustin, 2002). First, models were fit using all variables as input for both p1 and p2 in the two areas. Predictors were analyzed through the associated component smoothing functions (CSFs) to check for physical plausibility (Steger et al., 2016, 2021; Camera et al., 2021; Bajni et al., 2022, 2023). In addition, the rugs and confidence bands of the CSF graphs were analyzed to consider the possible introduction of variable cutoffs, to reduce model uncertainties due to scarce data with extreme values. Model fitting was then performed again with physically plausible variables only and cutoffs values. The penalization frequency of each variable was analyzed and only variables with CV penalization frequency lower than 75 % were kept in the final calculations. A variable was considered penalized for a component smoothing function with effective degrees of freedom lower than 0.7 (Bajni et al., 2023). Concurvity among independent variables was checked too, and in case of pairs of variables showing concurvity higher than 0.8, considered critical (Camera et al., 2021), the variable penalized most often than the other was excluded. Group 2 did not apply any rescaling nor standardization of the input variables.

Calibration–Validation approach. A non–spatial, k–fold CV (five folds, 20 repetitions) was carried out. To consider uncertainties, 100 instances of the model were fit with the selected variables using a random sample of 80 % of the available SUs in each study area. The 100 instances were applied to all SUs. Mean, median, and the difference between the 95th percentile and the 5th percentile susceptibility values were used as uncertainty. The remaining 20 % of the SUs were used for the optimized model validation.

Model evaluation. Model evaluation was performed based on AUC_{ROC}, and variable importance was checked calculating the mean decrease in explained deviance. This procedure was adopted for both step 1 and step 2, with the exception of the calibration–validation scheme that in step 2 was modified for comparability with results of other groups. In all figures and tables, results corresponding to this paragraph are labeled as GAM_1 (*cf.* Fig. 3).

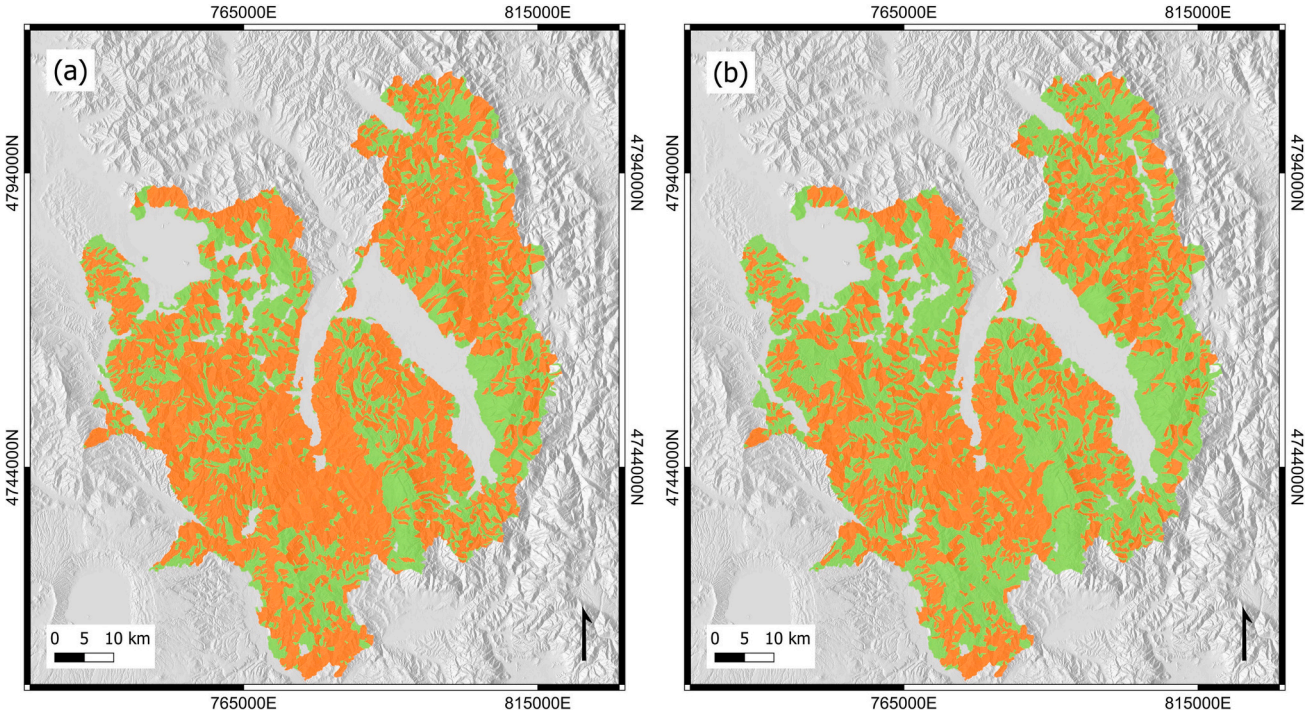
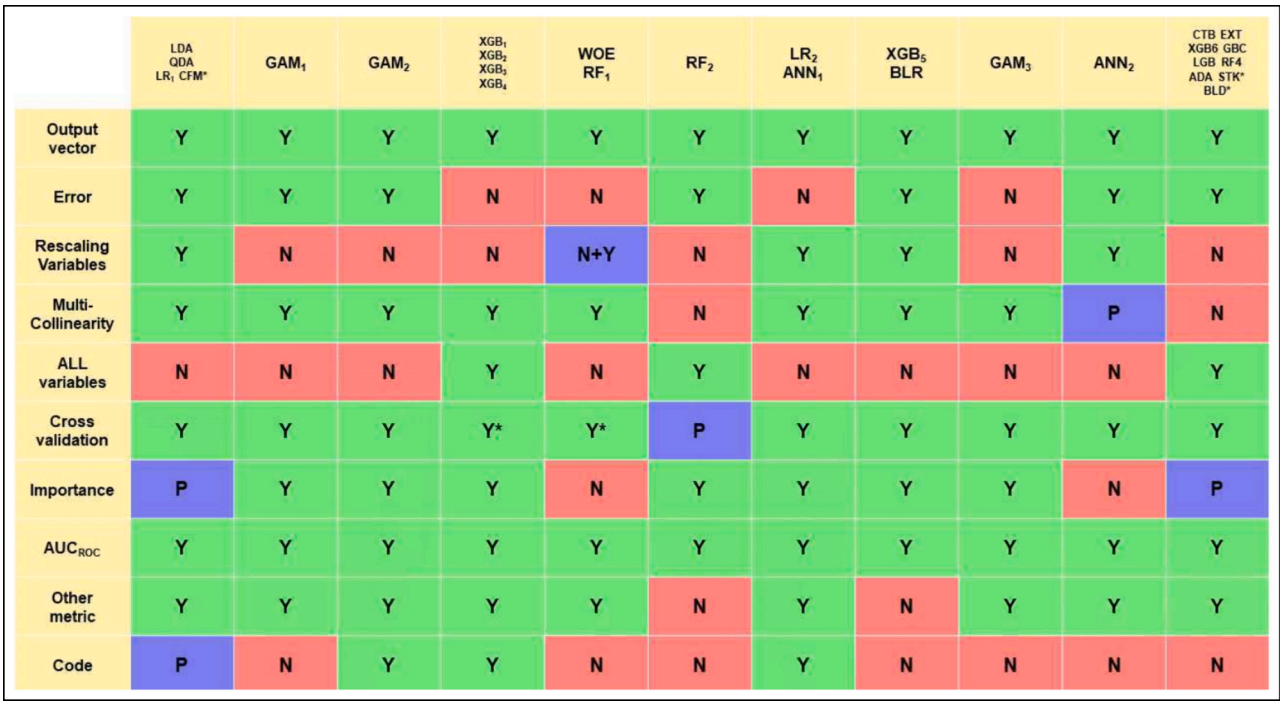


Fig. 2. Spatial distribution of positive (with landslides; orange) and negative (without landslides; green) slope units, in the dataset proposed in this work (*cf.* Fig. 1). Landslide presence is either from the field p1 (a) or p2 (b) in the attribute table (*cf.* Section 4, Tables 1 and 3). Background is a shaded relief map obtained from the European digital elevation model (EU–DEM). Maps are in EPSG:32632. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



	LDA QDA LR$_1$ CFM*	GAM$_1$	GAM$_2$	XGB$_1$ XGB$_2$ XGB$_3$ XGB$_4$	WOE RF$_1$	RF$_2$	LR$_2$ ANN$_1$	XGB$_5$ BLR	GAM$_3$	ANN$_2$	CTB EXT XGB6 GBC LGB RF4 ADA STK* BLD*
Output vector	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Error	Y	Y	Y	N	N	Y	N	Y	N	Y	Y
Rescaling Variables	Y	N	N	N	N+Y	N	Y	Y	N	Y	N
Multi-Collinearity	Y	Y	Y	Y	Y	N	Y	Y	Y	P	N
ALL variables	N	N	N	Y	N	Y	N	N	N	N	Y
Cross validation	Y	Y	Y	Y*	Y*	P	Y	Y	Y	Y	Y
Importance	P	Y	Y	Y	N	Y	Y	Y	Y	N	P
AUC$_{ROC}$	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Other metric	Y	Y	Y	Y	Y	N	Y	N	Y	Y	Y
Code	P	N	Y	Y	N	N	Y	N	N	N	N

Fig. 3. Results of the survey proposed to the participants of step one of the experiment, described in Section 4. Positive "Y," negative "N," partial "P" are the possible answers to the question of whether the features in each column were implemented; asterisks in CFM*, STK*, and BLD* denote that they were obtained as combination of other methods; asterisks in the "Cross–validation" row correspond to application of spatial CV.

4.3. Group 3. LSM via Bayesian generalized additive models

This contribution was presented as the EGU abstracts by Loche et al. (2023); Scaringi and Loche (2023).

Model selection. Group 3 (G3) used a GAM to investigate a specific research question related to the recently published national susceptibility maps of Italy (Loche et al., 2022). Because the proposed dataset is a subset of the data used at national scale, G3 compared the national LSM with the result for the smaller area, and evaluated the response of fixed and random effects, used in both works. For the within–sample step (calibration), G3 implemented a Bayesian version of a binomial GAM in R–INLA (Fang et al., 2023), which can explain the spatial distribution of landslides *via* a family of the Bernoulli exponential function (Lindgren and Rue, 2015). This framework allowed G3 to model fixed effects as linearities and random effects as nonlinearities, and to assess their associated uncertainty. The fitting procedure returned satisfactory results (acceptable or excellent performances) in terms of AUC$_{ROC}$, with values around 0.8 (Hosmer Jr. et al., 2013).

Variable selection. The model performance is affected by the number of variables, which was intentionally kept low for ease of interpretation and to keep calculations as simple as possible (Lombardo and Mai, 2018). Group 3 used the `corrplot` R package (Wei and Simko, 2021) to explore multicollinearity issues (Allen, 1997). They computed a coefficient for each of the independent variables and produced a graphical display of a correlation matrix, regressing all the independent variables against each other. As a general rule, multicollinearity is a potential problem when the coefficient is higher than 0.75, and a serious problem when it is higher than 0.9 (Mela and Kopalle, 2002). Based on this analysis, excluded variables were `slope_stdd`, `tcurv_aver`, `tcurv_stdd`, and `BDRICM_std` (refer to Table 1). Instead, `elev_aver` and `elev_stdd` have not been considered *a priori*, following the original setting of Loche et al. (2022). Group 3 did not apply any rescaling nor standardization of the input variables.

Calibration–Validation approach. The within–sample test (calibration) is described in the paragraph 'Model selection.' Mirroring the goodness–of–fit assessment, G3 also evaluated the out–of–sample performance. They performed a ten–fold CV with mutual exclusion to guarantee that no influence from repeated samples would affect the validation replicates. The variability resulting from the repetitions did not compromise the model output, for both landslide presence scenarios.

Model evaluation. Model evaluation was performed based on AUC$_{ROC}$. In all figures and tables, results corresponding to this paragraph are labeled as GAM$_2$ (*cf.* Fig. 3).

4.4. Group 4. Effect of cross–validation within the XGBoost method

This contribution was presented as the EGU abstract by Samodra et al. (2023).

Model selection. The XGBoost algorithm integrates multiple classification and regression trees (CARTs) and successively combines the output of weak learners to improve performance (Chen and Guestrin, 2016). The `Tidymodels` collection of R packages (Kuhn and Wickham, 2020) by Kuhn and Silge (2022) was used by Group 4 (G4) to execute the XGBoost algorithm.

Variable selection. Each SU has information about 26 controlling factors, in which all factors were used in the landslide susceptibility modeling processes using the XGBoost algorithm. Group 4 did not apply any rescaling nor standardization of the input variables.

Calibration–Validation approach. The XGBoost model with spatial and non–spatial CV was used to estimate the performance and the accuracy of landslide susceptibility model based on SUs. Group 4 split data into 75 % (5519 SU) for training and 25 % (1841 SU) for testing/success rate slope. The training dataset was used for CV and hyperparameter tuning, and the test data was set aside for independent validation.

A model without CV (XGB$_1$) was also used to show the existence of overfitting. Non–spatial CV (XGB$_2$) applied the random 10–fold CV, in which the samples were partitioned randomly into 10 folds of roughly equal size. Spatial CV was applied by partitioning 10–fold data spatially based on block CV (XGB$_3$) and clustering CV (XGB$_4$). The grid Latin hypercube implemented in `Tidymodels` was applied to tune both non–spatial and spatial CV strategies. Best tuning was automatically selected to evaluate the performance of XGBoost model applied to the

training set. Hyperparameter values from best tuning were selected to create the best model and to obtain the performance of XGBoost model represented by AUC_{ROC}.

Model evaluation. The performance of the model was assessed by AUC_{ROC}. In all figures and tables, results corresponding to this paragraph are labeled as $XGB_{1...4}$ (*cf.* Fig. 3), corresponding to no CV, non–spatial CV, spatial block CV, and spatial clustering CV, respectively.

4.5. Group 5. Weight of evidence versus random forest methods for LSMs

This contribution was presented as the EGU abstract by Sinčić et al. (2023).

Model selection. Group 5 (G5) selected the WoE and RF methods. The two methods were chosen as representative of two opposite approaches because WoE is a bivariate method used since early works on landslide susceptibility assessments and RF is a recent machine learning algorithm already commonly used (Reichenbach et al., 2018). Moreover, WoE requires only specifying unstable areas, whereas RF also requires knowledge of stable areas. Calculations for WoE method were performed in ArcMap 10.8.1 using the "Spatial Analyst Toolbox" and Microsoft Excel, whereas the "Statistics and Machine Learning Toolbox" in MATLAB software (version 9.10.0.1602886) was used for RF method.

Variable selection. Predictors were reclassified using natural breaks in ArcMap 10.8.1 into 10 classes, followed by testing collinearity using LAND–SUITE software (Rossi et al., 2022), which showed a significant number of predictors having a Pearson *R* absolute value of 0.5 or greater. This resulted in keeping 11 variables, namely: `slope_stdd`, `pcurv_aver`, `nthns_aver`, `nthns_stdd`, `easns_aver`, `easns_stdd`, `twi_stdd`, `DBRICM_ave`, `CLYPPT_ave`, `CLYPPT_std`, `Max_Distan` (refer to Table 1).

Calibration–Validation approach. To apply the two methods on similar footing, the same unstable SUs were selected for both WoE and RF, separately in the `p1` and `p2` scenarios. Thus, 50 % of available unstable SUs were selected for training the model, *i.e.*, 1797 in `p1` and 1136 in `p2`. The selection of an equal number of stable SUs for the RF method was done with the assumption that they can be anywhere, except the 50 % of unstable SUs already selected for training. In other words, stable SUs were selected from the area which had excluded only the previously selected unstable SUs, assuming that information about stable SUs in the original dataset is unknown. As a result, 573 stable SUs in `p1` and 211 stable SUs in `p2` scenarios for RF model training were flagged as unstable in the original dataset, thus overriding the original classification. The latter resulted in a skewed landslide datasets for training RF but ensured an unbiased landslide sampling procedure for stable SUs. The WoE method required no pre–processing, whereas the predictors were normalized for the RF method.

Model evaluation. For fitting performance, 50 % of unstable SUs selected for implementing the methods were examined, whereas for predictive performance the remaining 50 % of unstable SUs were used. The AUC_{ROC} values were defined with cumulative percentage of study area in susceptibility classes and the cumulative percentage of landslide area in susceptibility classes. The latter resulted in success and prediction rates for fitting and predictive performance, respectively (Chung and Fabbri, 1999, 2003). On the other hand, all unstable and all stable SUs were used to define a hit rate and false alarm rate curve for which an AUC_{ROC} was calculated. Moreover, overall accuracy was determined at the 0.5 threshold for all four LSMs. To compare the approach of two different methods where WoE uses only unstable SUs and RF uses both unstable and stable SUs, the fitting and predictive performance were measured by observing only unstable SUs, whereas classification parameters present additional metrics to measure the LSMs considering both stable and unstable SUs. Model evaluation calculations in all cases were performed in ArcMap 10.8.1 using the Spatial Analyst Toolbox and Microsoft Excel software. The focus of the described approach was on modeling properties including different methods, stable and unstable SU sampling, and the two presence flags differences. In all figures and tables, results corresponding to this paragraph are labeled as WOE and RF_1 (*cf.* Fig. 3).

4.6. Group 6. Random forest as a high accuracy model for LSM

This contribution was presented as the EGU abstract by Sirbu (2023).

Model selection. Group 6 (G6) selected RF, a regression and classification algorithm that uses multiple CART to produce high–accuracy models (Breiman, 2011). Random forest is non–parametric (Merghadi et al., 2020) and is able to handle both linear and nonlinear processes. Thus, RF is increasingly used in landslide modeling (Reichenbach et al., 2018; Zeng et al., 2023), and it often outperforms other statistical and machine learning algorithms (Goetz et al., 2015). In this work, the model was set up as a script in R software (R Core Team, 2021) using the following packages: (i) `randomForest` (Liaw and Wiener, 2002) to run the algorithm, (ii) `ROCR` (Sing et al., 2005) to evaluate the performance of the model and to run the validation for the results, and (iii) `rgdal` (Bivand et al., 2023) to read input data and to produce the output. The only parameter of the model was the number of CARTs, which was set using $ntree = 1,501$.

Variable selection. The model set up by G6 produced a ranking of the input variables based on two algorithms, mean decrease in accuracy and the decrease in node impurity (using the Gini index), for each presence scenario. Ranking means neglecting one predictor, assessing the accuracy of the model and Gini index, and repeating for every predictor. If the model has a high accuracy without one predictor, that is considered less important. The model was trained using all of the 26 independent variables and 7360 SUs. Because the method performs multiple classifications based on CARTs, the results are robust and outliers have little relevance; thus a multicollinearity test of the variables was not essential (Lee et al., 2018). Group 6 did not apply any rescaling nor standardization of the input variables.

Calibration–Validation approach. The input data was split 70 % and 30 % into training and validation data, respectively. The first split was used to train the model, to assess the model settings (*e.g.*, *ntree* parameter), and to compute the input variable ranking.

Model evaluation. The validation data were used to assess the accuracy of the model using the AUC_{ROC} metric for each of the two presence scenarios. In all figures and tables, results corresponding to this paragraph are labeled as RF_2 (*cf.* Fig. 3).

4.7. Group 7. LSM with logistic regression and artificial neural networks

This contribution was presented as the EGU abstract by Torizin and Schüßler (2023).

Model selection. Following an initial in–depth exploration of the preliminary dataset, which revealed significant correlations among covariates and poor bivariate separability of landslide presence labels with a specific single covariate, Group 7 (G7) decided to use multivariate methods. Their choice encompassed a linear model utilizing LR (*e.g.*, Steger et al. (2016); Lombardo and Mai (2018)) and an ANN in the shape of a multi–layer perceptron (MLP; Ivakhnenko and Lapa (1967)). The latter should uncover possible non–linear effects and increase the separability in multivariate cases. Logistic regression was trained on the labeled dataset using stochastic gradient descent to minimize the loss given by the binary cross entropy.

An ANN is trained by adjusting the connections' weights between neurons through error backpropagation (Rumelhart et al., 1986). Group 7 used a relatively simple ANN with one hidden layer (Ermini et al., 2005; Lee and Evangelista, 2006), with 100 neurons and one output layer, giving the probability for the target label to be unity. The neurons in the hidden layer use rectified linear unit (ReLU) as an activation function to handle possible non–linear relations in the data. Group 7 used binary cross entropy as a loss function, and Adam (which stands for adaptive moment estimation) as a solver (Kingma and Ba, 2017).

To build the models and conduct the analysis, G7 harnessed the

powerful `Scikit–Learn` Python library (Pedregosa et al., 2011) to perform the analysis, capitalizing on its slim coding and high efficiency for building machine–learning models.

Variable selection. To assess the predictive power of each variable, G7 conducted individual bivariate modeling, evaluating the performance using the preliminary dataset. They proceeded with the most influential predictor and incrementally added covariates stepwise. Model performance was assessed at each step, removing the last added covariate if the AUC_{ROC} did not increase. The stepwise approach resulted in fewer predictors, with 9 to 10 for LR and 10 to 14 for ANN depending on the used presence label, exhibiting robust performance on the test data for both models. Additionally, predictors were scaled to a range of [0,1] using min–max normalization.

Calibration–Validation approach. The data preparation phase involved randomly splitting the dataset into training and test sets using a 70:30 ratio for CV.

Model evaluation. To evaluate the performance, G7 utilized AUC_{ROC} and success rate (Chung and Fabbri, 2003). In all figures and tables, results corresponding to this paragraph are labeled as LR_2 and ANN_1 (*cf.* Fig. 3).

4.8. Group 8. Bayesian logistic regression and optimized XGBoost models for LSMs

This contribution was presented as the EGU abstract by Mirus and Woodard (2023).

Model selection. Group 8 (G8) used two methods, XGBoost (Chen and Guestrin, 2016) and a Bayesian implementation of logistic regression (BLR). XGBoost is fast, straightforward to implement, and has produced accurate susceptibility maps at other locations (Sahin, 2020). Group 8 optimized the model hyperparameters suggested by the model developers (*i.e.*, 'max_depth,' 'min_child_weight,' 'subsample,' 'gamma,' and 'colsample_bytree') using a Bayesian CV procedure. Logistic regression is the most used algorithm for susceptibility analysis (Reichenbach et al., 2018) and its Bayesian implementation allowed G8 to account for uncertainty in the estimated model coefficients.

Variable selection. For each of the two methods, G8 generated models with three groups of input data for each target variable (six datasets total): (1) all the available predictors except `area`; (2) only `slope_aver` and `slope_stdd`; (3) all predictors with a variance inflation factor (VIF, a measure of collinearity between variables within a model) less than five, which is a conservative value (James et al., 2013).

For each data group, G8 first standardized the predictors to have a mean of zero and standard deviation of one to increase computational efficiency and to better constrain the coefficient priors for the BLR method. They measured VIF values of the predictors using an iterative approach. Specifically, they generated a frequentist LR model, measured the VIF of the predictor variables, eliminated the highest tenth percentile of variables with VIF values greater than five, and repeated the process until all predictors had a VIF value less than five. This implementation allowed G8 to account for differences in the measured VIF values from different predictor combinations. Using this approach, `slope_aver`, `tcurv_stdd`, `BDRICM_std` and `SNDPPT_ave` (refer to Table 1) were excluded in the `p1` target variable dataset and `slope_aver`, `twi_averag`, `BDRICM_std` and `SNDPPT_ave` were excluded from the `p2` target variable dataset. Group 8 standardized variables to have a mean of zero and standard deviation of one.

Calibration–Validation approach. Group 8 used k–fold CV (10 folds and 10 repeats) with XGBoost after optimizing the model hyperparameters with half of the available data, and used the LOO CV technique (70:30 random split) with the LR model. The chosen CV techniques provided uncertainties in model performance for both algorithms.

Model evaluation. Group 8 measured model performance using AUC_{ROC} and the Brier score (B, mean–square error). Brier score is defined as follows:

$$B = \frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2, \tag{1}$$

where P is the model prediction (*i.e.*, probability), O is a binary variable (here, landslide presence), and N is the number of observations (Brier, 1950). In all figures and tables, results corresponding to this paragraph are labeled as XGB_5 and BLR (*cf.* Fig. 3).

4.9. Group 9. Exploring the role of slope unit size in LSM using GAMs

This contribution was presented as the EGU abstract by Moreno and Steger (2023).

Model selection. The approach of Group 9 (G9) leveraged GAMs to address the challenge of spatially varying SU sizes in landslide susceptibility modeling. Generalized additive models are flexible extensions of the well–known generalized linear models that allow accounting for non–linear relationships between predictors and the target (Zuur et al., 2009). The approach builds upon the comprehensive R package `mgcv` (Wood, 2017).

The SU–based approach for LSMs has recently garnered substantial attention for their flexibility in accommodating diverse responses, encompassing binary output (presence/absence), and count (number of landslides), and reduced sensitivity to inaccuracies in landslide positional referencing. However, a pivotal aspect to underscore is the inherent variability in SU sizes across a study area, potentially resulting in spatially varying likelihoods of SUs being affected by landslides. Group 9 assumed that larger SUs are not necessarily more susceptible to landslides but are more likely to be categorized as unstable. This methodological aspect may affect the subsequent susceptibility models, especially if predictors correlate with SU size.

Variable selection. Group 9 formulated four distinct strategies described below. The first strategy (Model 1) involved all predictors, including the SU area, in the model fitting and prediction. Building upon this setup, the second strategy (Model 2) maintained the same suite of predictors while explicitly ignoring the SU area in model fitting and spatial prediction. The third strategy (Model 3) was devised to showcase the discriminatory capacity of SU size in discerning SUs with landslides from those without, relying on a single–variable model that hinges solely on the SU area as a predictor. Finally, the fourth strategy (Model 4) used all predictors during model fitting but excluding the effects of SU size from spatial prediction. This is achieved by setting to zero the smooth component of the SU size; in other words, the model was allowed to learn from the SU size, but its explained variability was zeroed during the predictions. In that way, G9 ensured that the effect of SU size and its potential confounding effects were contained within the model fitting process but not reproduced directly or indirectly into the predictions.

Calibration–Validation approach. The performance of the models was evaluated through 10–fold random CV and 10–fold spatial CV, each conducted with ten repetitions in the R package `sperrorest` (Brenning, 2012). The analysis and comparison of the four strategies started with exploratory analyses, which involved correlation plots and relative variable importance assessments. These analytical procedures were carried out to identify potential correlations among predictors in a straightforward manner. Subsequently, the four models were fit and initially assessed through the interpretation of partial effect plots, and their dependency on the SU area was examined *via* scatterplots. As a last step, G9 visually compared the four resulting susceptibility models and discussed the benefits and limitations, highlighting the proposed solution (Model 4).

Model evaluation. Model evaluation was performed based on AUC_{ROC}. Here, we only report performance results for Model 4. In all figures and tables, results corresponding to this paragraph are labeled as GAM_3 (*cf.* Fig. 3).

4.10. Group 10. Role of feature selection on the prediction performance of a neural network–based LSM

This contribution was presented as the EGU abstract by Satyam et al. (2023).

Model selection. Group 10 (G10) used an RF and an MLP, a variety of ANN, to develop LSMs using the selected features. The choice of model was also based on trial and error; G10 applied both models for the same dataset. They used RandomForestClassifier and MPLClassifier from the Python package `Scikit–learn` (Pedregosa et al., 2011). Even though RF provided better performance metrics, it was not chosen for the final analysis, considering the computational time taken for optimizing the hyperparameters. Group 10 used the `GridsearchCV` function (Pedregosa et al., 2011) to tune the hyperparameters of the MLP. The hidden layer size, initial learning rate, and the activation function were selected using `GridsearchCV`, for each set of input variables.

Variable selection. Group 10 used all variables except SU area and checked the correlation matrix and VIF. They found that several variables were highly correlated with each other, and there were VIF values as high as 535.19 in the dataset. From this observation, they decided not to consider all variables for the analysis. Group 10 used the `SelectKBest` function (Pedregosa et al., 2011), a feature selection function based on univariate statistical tests. All of the predictors were scaled to a range of [0,1] using min–max normalization.

Calibration–Validation approach. The benchmark dataset was divided into 80 % data for training and testing using a 5–fold CV, and the remaining 20 % was kept for independent validation. The first 80 % of data was used to select the *K* "best" variables. In this study, G10 varied the number of 'best variables' from 5 to 15, to understand the effect of the number of variables. After selecting the variables, the data were updated with only the selected variables and the target variable (`p1` or `p2`).

Model evaluation. Accuracy and AUC_{ROC} were used to compare the performance of different outputs. The trained model was then used to predict the probability of occurrence of landslides on the validation dataset, and the accuracy and AUC_{ROC} values of this dataset were used for comparing the performances. It was observed that the AUC_{ROC} values increased with a larger number of variables. However, the VIF values were too high in such cases, and again by trial and error, G10 limited the number of variables to 10, searching for a balance between importance of variables and AUC_{ROC} values. In all figures and tables, results corresponding to this paragraph are labeled as ANN_2 (*cf.* Fig. 3).

4.11. Group 11. Ensemble learning with spatial cross–validation

This contribution was presented as the EGU abstract by Aguilera et al. (2023).

Model selection. Group 11 (G11) used an ensemble machine learning approach to enhance the performance and generalization ability of the LSM using the preliminary dataset. They utilized various ensemble techniques, including bagging, boosting, stacking, and blending. Bagging techniques, specifically RF and extremely randomized trees (EXT), trained independent models on different bootstrap samples of the training data. The predictions from these models were combined through voting, resulting in a strong and accurate model. Random forest utilized random feature selection to increase diversity and reduce overfitting, and EXT further randomized feature selection and splitting thresholds, potentially improving computational efficiency.

Boosting techniques, such as Gradient Boosting Classifier (GBC), Extreme Gradient Boosting (XGB), Light Gradient–Boosting Machine (LightGBM), CatBoost (CBT), and AdaBoost (ADA), were aimed at building powerful models by iteratively training weak models and emphasizing misclassified instances. The final predictions were obtained through weighted voting, reducing bias, and improving overall accuracy. Each boosting algorithm used unique optimization strategies and offered distinct advantages.

Stacking (STK) involved training a meta–model on out–of–fold predictions from different base models during k–fold CV. This technique aims to leverage the strengths of different models and achieve improved performance.

Blending (BLD), a simplified version of STK, directly combines predictions from multiple models without any meta–model. The final predictions were obtained using soft voting, which involves summing the predicted probabilities for class labels and predicting the class label with the largest sum probability.

Variable selection. Group 11 implemented their calculations as a script in Python v3 using the following packages: (i) for data analysis and manipulation, they used `NumPy` (Harris et al., 2020), `Pandas` (McKinney, 2010), and `GeoPandas` (Jordahl et al., 2020); (ii) to develop the machine learning models, they utilized `PyCaret` (Ali, 2020), an `AutoML` package based on `Scikit–learn` (Pedregosa et al., 2011), which allows integration with many other packages. Group 11 conducted experiments with these ensemble methods using the raw dataset, excluding `area`, and did not apply any rescaling or standardization of the input variables.

Calibration–Validation approach. Group 11 applied spatial CV using spatial blocks. Spatial CV is particularly relevant where spatial autocorrelation exists in the training data, such as clustering of data points in space (Beigaitė et al., 2022; Meyer et al., 2019; Schratz et al., 2019). This approach prioritizes consistency over accuracy. In the case of the benchmark dataset, the SUs in the landslide susceptibility maps exhibited spatial patterns. For each method, G11 used 10–fold spatial CV and averaged the results of 100 runs from the best–performing classifiers to estimate uncertainty.

Model evaluation. Model performance evaluation was based on the different classification metrics (AUC_{ROC}, accuracy, precision, recall, F1 score, and Kappa coefficient). In all figures and tables, results corresponding to this paragraph are labeled as CTB, EXT, XGB_6, GBC, RF_3, ADA, STK, and BLD (*cf.* Fig. 3).

Table 2
A summary of the models applied by all participant groups in step one of the experiment, described in detail in Sections 4.1–4.11. The columns `p1` and `p2` describe whether each group calculated results for the corresponding landslide presence scenario.

Model	Group	p1	p2	Description
LDA	G1	x	x	Linear discriminant analysis
LR_1	G1	x	x	Logistic regression
QDA	G1	x	x	Quadratic discriminant analysis
CFM	G1	x	x	Combined forecast model
GAM_1	G2	x	x	Generalized additive model
GAM_2	G3	x	x	Generalized additive model
XGB_1	G4	x		XGBoost, no cross validation (CV)
XGB_2	G4	x		XGBoost, non–spatial 10–fold CV
XGB_3	G4	x		XGBoost, spatial block CV
XGB_4	G4	x		XGBoost, spatial clustering CV
WoE	G5	x	x	Weight of evidence
RF_1	G5	x	x	Random forest
RF_2	G6	x	x	Random forest
LR_2	G7	x	x	Logistic regression
ANN_1	G7	x	x	Artificial neural network
XGB_5	G8	x	x	XGBoost
BLR	G8	x	x	Bayesian logistic regression
GAM_3	G9	x		Generalized additive model
ANN_2	G10	x	x	Artificial neural network
CTB	G11	x	x	CatBoost
EXT	G11	x	x	Extreme gradient boosting
XGB_6	G11	x	x	XGBoost
GBC	G11	x	x	Gradient boosting classifier
RF_3	G11	x	x	Random forest
ADA	G11	x	x	AdaBoost
STK	G11	x	x	Restaker linear regression (combined model)
BLD	G11	x	x	Blender (combined model)

4.12. Overview and results of methods chosen by contributors in step one

The above sections outlined a plethora of approaches and choices intrinsic to an LSM exercise. Fig. 3 shows answers to specific questions, provided by the contributors along with results of their calculations for step one of the experiment. One can observe that the only features common to all of the groups were to produce a vector output and calculate AUC_{ROC} as a performance metric. Many groups also provided an uncertainty along with their results, most of them checked for collinearity and dropped a few variables, using only a subset of the initial variables for their classification, with different mixtures. Most of the contributors performed the frequently used training–validation data split, and a few implemented spatial cross validation.

Table 2 lists all the models considered at step one of this experiment, including model names, group who applied the models, whether they calculated results for `p1` and/or `p2`, and a one–line description of the model.

Overall, we identified a few key points in which these approaches differ, and that can potentially affect the susceptibility values and model performance. Firstly, both statistical (*e.g.*, LR, LDA, QDA, and WoE) and machine learning models (*e.g.*, ANN, RF, and XGBoost) were used. Secondly, a few authors have included all variables, and others have used analytical or heuristic approaches prior to – or during the modeling exercise – to remove some variables. Thirdly, the approaches differed in the way data were split to calibrate and validate the model. Although all authors performed separate calibration and validation with variable fractions of input data, the use of CV and spatial CV was not systematic. In particular, only G4 and G11 applied spatial CV.

A few groups selected more than one method, two groups (G1 and G11) proposed ensemble modeling with different ways of combining a few results into a single one. A few groups selected the same method: LR and Bayesian LR were used by G1, G7, and G8; generalized additive models were adopted by G2, G3, and G9; a tree boosting system called XGBoost was used by G4, G8, and G11; RF was adopted by G5, G6, and G11.

Although many contributing teams chose to remove some variables, they differed in how they decided to do so, ranging from pre–modeling heuristics (*e.g.*, based on a VIF–thresholding, G8 and G10), and penalization–based methods during modeling (*e.g.*, G2). Likewise, although many groups performed a CV, they differed in the number of folds/repetitions.

As expected, all of the participants used AUC_{ROC} as a performance metric. In addition, G1 considered explicitly the graphical representation of four–fold plots (not shown here); G5 and G7 utilized success/prediction rate to evaluate performance on the basis of SU ranking of the results of classification; G8 suggested the use of Brier score, equivalent to the mean of squared differences between the (probabilistic) prediction and the target variable in each SU; G11 used several performance metrics, including accuracy, precision, F1, and Cohen's Kappa.

Due to the large number of methods applied in the first part of the experiment, and the heterogeneity of the application approaches, we do not show susceptibility maps in this section. We will show maps corresponding to the final benchmark dataset devised in this experiment, instead, described in the next section.

Fig. 4 shows a pairwise comparison of results of step one, calculated as follows:

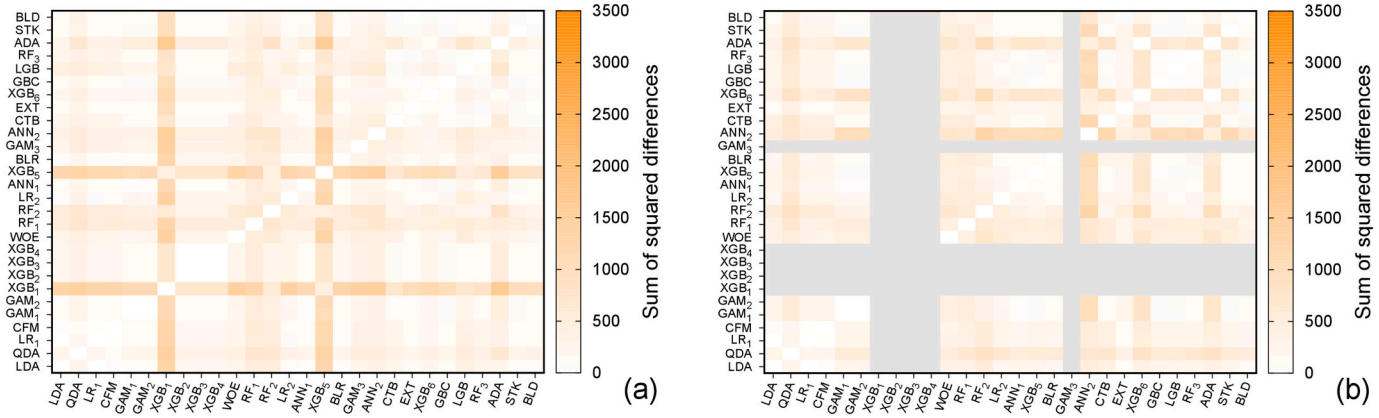


Fig. 4. Pairwise comparison of the results of different methods, in step one of the experiment, described in Section 4, calculated as in Eq. (2). Panels (a) and (b) correspond to the target variables `p1` and `p2`, respectively. Names of the different methods are as in Fig. 3, Sections 4.1–4.11. Grey color denotes missing data.

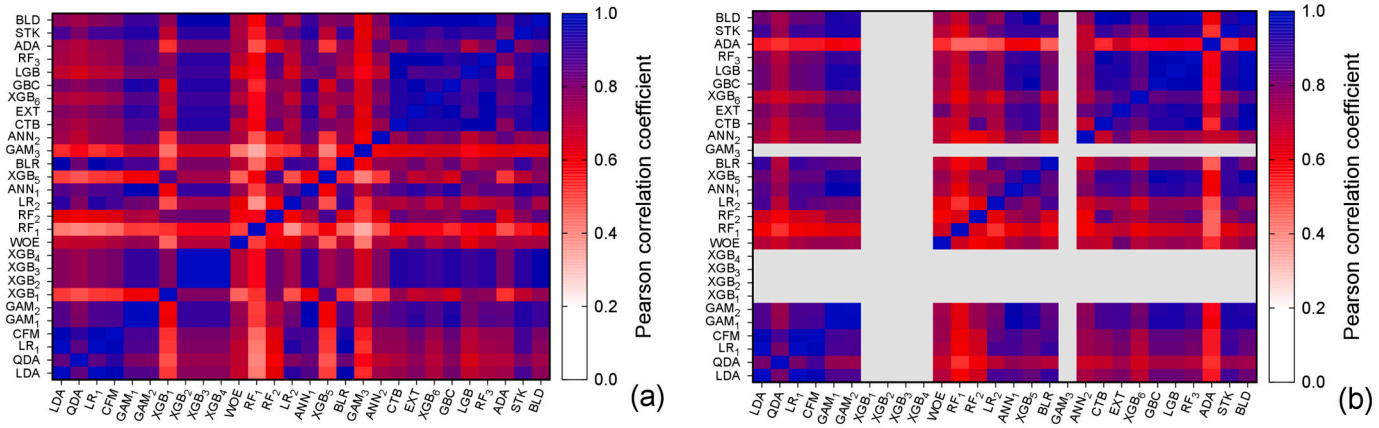


Fig. 5. As in Fig. 4, but for pairwise correlations between results for different methods. Grey color denotes missing data.

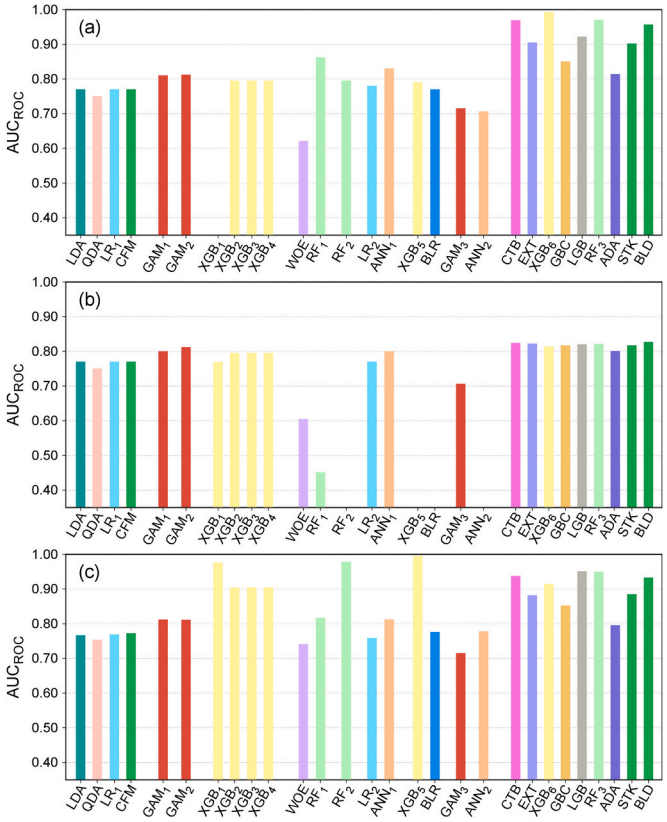


Fig. 6. Comparison of area under the receiver operating characteristic (AUC_{ROC}) values for the different methods considered in step one of the experiment, limited to the `p1` landslide flag. Both in calibration (a) and in validation (b) the same method gives different results, due to the different workflow of application by different research groups. The plot in (c) was obtained by the organizers of the experiment, calculating AUC_{ROC} from the numerical results provided by contributors. Names of the different methods are as in Fig. 3, Sections 4.1–4.11.

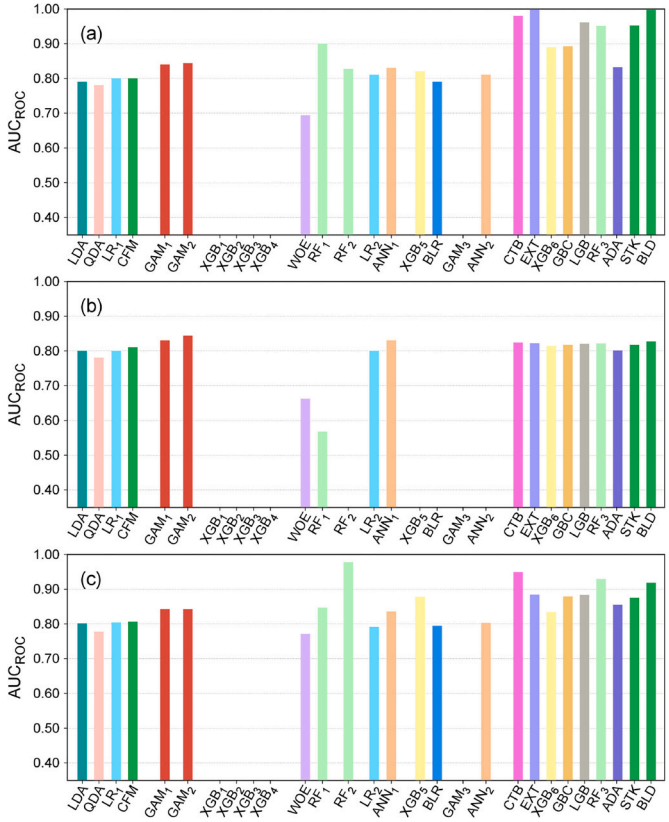


Fig. 7. As in Fig. 6, but for the `p2` landslide presence flag.

$$D_{ij} = \sum_{k=1}^{N_{SU}} \left(S_i^k - S_j^k \right)^2, \tag{2}$$

where i and j run in the set of methods, k labels the SUs in the dataset, and $0 \leq S_i^k \leq 1$ is the susceptibility value of the i–th SU, calculated with the k–th method. One can observe from Fig. 4 that adoption of the same method did not necessarily lead to very similar results. For example, the largest differences, according to the criterion of Eq. (2), are between XGB_1, XGB_5 and all other methods; the difference between RF_1, RF_2, ADA and other methods stands out as well (Fig. 4(a), corresponding to `p1`). For the target label `p2`, Fig. 4(b), the classification from $XGB_{1...4}$ gave the (exactly) same results in this case, and for `p1` XGB_1 differs slightly from the $XGB_{2..4}$: they differ from the former only for the spatial CV strategy which, thus, seems to be marginally relevant, here. A few values are missing in the figures because no results were provided by the participants at step one. On the other hand, we note that results for the GAM method (`p1`) in the variants GAM_1 and GAM_2 are similar, although they somewhat differ from the variant GAM_3. The reason for that may reside in the different way of using predictors in the application of GAM_3, in which the role of SU size (`area`) was emphasized.

In both panels of Fig. 4 (for `p1`–`p2`), the results by G1 (LDA, QDA, LRM, and CFM with the software LAND–SUITE) and by G11 (CTB, EXT, XGB_6, GBC, LGB, RF_3, ADA, STK, and BLD with an ensemble machine learning) show similarities within each subset, indicating that different methods applied by the same author somehow produce more similar results. This may reside in pre–processing of data, variable selection, and training–validation strategy. We reached the same conclusion looking at Fig. 5, which shows pairwise correlations between results, calculated with the `cor()` function of the `corrplot` package in R (Wei and Simko, 2021).

Fig. 5 also shows a few high–correlation blocks (*e.g.*, the blueish blocks between methods GAM_1, GAM_2, and $XGB_{2..4}$; the methods within the ensemble machine learning by G11 for the `p1` case; and similar cases for `p2`). Interpretation of this occurrence is not straightforward, because they correspond to multiple methods applied by different authors. Results for $XGB_{1..4}$ and GAM_3 were not provided for `p2` (grey bands in the figures).

Figs. 6 and 7 show AUC_{ROC} values for all the different methods, for training and validation reported by the authors. Moreover, we obtained AUC_{ROC} independently, from susceptibility values in the attribute tables provided by the users and the target values `p1`–`p2`, using the `roc()` function of the `pROC` R package (Robin et al., 2011). In the figures, a few missing entries are due to a few authors providing only results for `p1` and/or only for the training step. Colors are consistent between the same method applied by different groups; combined models (CFM, STK, and BLD) also share the same color (dark green).

A general comment about AUC_{ROC} results is that the validation values are systematically lower than the training (fit) values, which is by design. We observe that the calculated performance is often different from the values provided by the users, both in excess or in deficiency.

This is also partially expected because most authors classified the final maps using a different subsample of the provided dataset. However, the $XGB_{1...4}$, WOE, RF_2, and XGB_5 (p1) models resulted in higher AUC_{ROC} measured by the organizers than both training and validation values, denoting some effect on how the final maps were assembled, possibly other than a combination of training (fit) and validation (predicted) results.

Lower values for the validation case are more prominent for a few methods. They are prominent in RF_1, for both landslide presence scenarios (likely due to the stable SU sampling strategy), and in CTB, XGB_6, RF_3, and BLD, to a lesser degree. This indicates potential overfitting of the training data and diminished performance on unseen data. However, ensemble methods from G11 (including CTB, XGB_6, RF3, and BLD) exhibit the highest validation scores among all methods (Figs. 6 and 7). Thus, the results obtained from the spatial CV scheme appear consistent and generalizable in these cases, despite the slight overfitting observed in the models.

The findings described in this section allowed us to draw the conclusion that a truly useful benchmark dataset for LSMs would be complemented by a minimal and well–defined set of prescriptions about application of the classification methods. As a result, the variability in the output LSMs include a substantial component due to such choices. In the next section, we describe both changes in the proposed dataset and a set of prescriptions for such choices, aimed at minimizing the effect of methodological choices to obtain a meaningful benchmark.

5. Final assessment of the benchmark dataset for landslide susceptibility

The selection of input variables – along with the type of model applied – has a large effect on the final susceptibility results. Thus, to improve the comparability of the models, we introduced a second step (step two) whereby we updated the input variables and asked the contributing teams to include the entire, updated dataset in their susceptibility model.

We updated the dataset firstly by including lithological information from the geo–mechanical lithology map of Italy by Bucci et al. (2022) because many contributors asked to include such information. We first calculated the percentage presence of lithological classes in the whole area, and selected the five classes with largest percentage cover, namely: alluvial deposits (*Al*, 12 %), unconsolidated sedimentary rocks (*Ucr*, 27 %), marlstone (*M*, 4 %), schistose metamorphic rocks (*Ssr*, 35 %), and carbonate rocks (*Cr*, 18 %). Total percentage was 96 %. Fig. 8 shows a simple description of the new variables. Lithological classes were provided as areal percentage in each SU polygon, which includes information about SUs containing different lithologies.

Fully comparable results also required each participant to adopt the same workflow concerning training/validation steps. To do that, the organizers had contributors apply a 10–fold CV procedure with mutual exclusion. That amounts to splitting the dataset into 10 numerically balanced parts. Training would be applied 10 times, on 90 % of the data, and validation would be performed on the remaining 10 %. Iterating the procedure 10 times on the 10 different splits for validation provides a fully validated LSM. The advantage of this procedure is that susceptibility values in each SU are calculated with a model trained with independent data.

5.1. Removing correlations

The updated dataset contains the following variables: slope, curvatures (morphometric) northerness/easterness, elevation, TWI (topographic wetness index), max distance and $MD/\sqrt{A}$ (max distance over the square root of SU area), bulk density, clay/sand/silt content (related to soil properties and texture), and percentage of lithology classes.

A few of these quantities are probably strongly correlated, and we

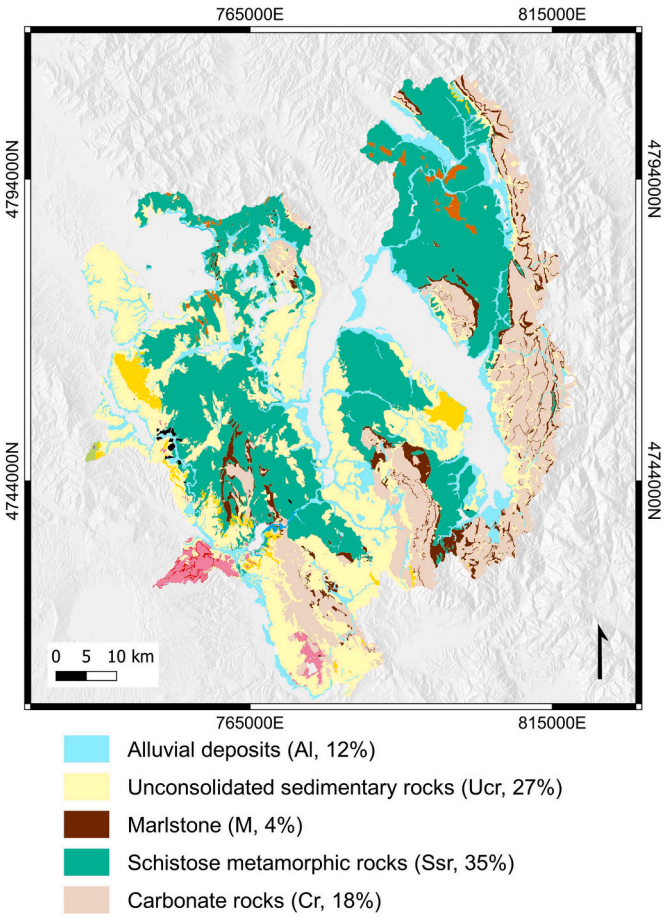


Fig. 8. Lithological classes included in the final version of the benchmark dataset. The figure shows a subset of the map prepared for the whole of Italy by Bucci et al. (2022). For this study, we considered only the five most representative classes as predictors, covering 96 % of the study area. Shaded relief map as in Fig. 2.

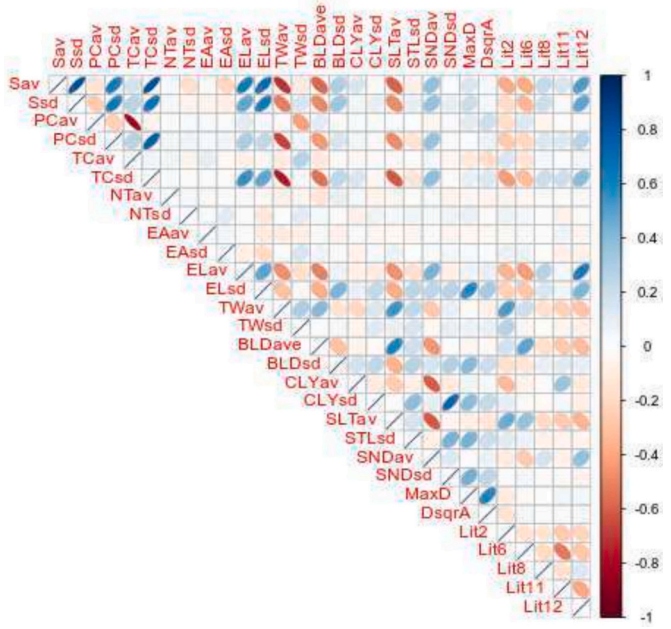


Fig. 9. Linear correlation plot of the 29 candidate variables for the final benchmark dataset. Short names of variables are as in Table 1. Symbols with larger ellipticity correspond to a higher degree of correlation between the two considered variables.

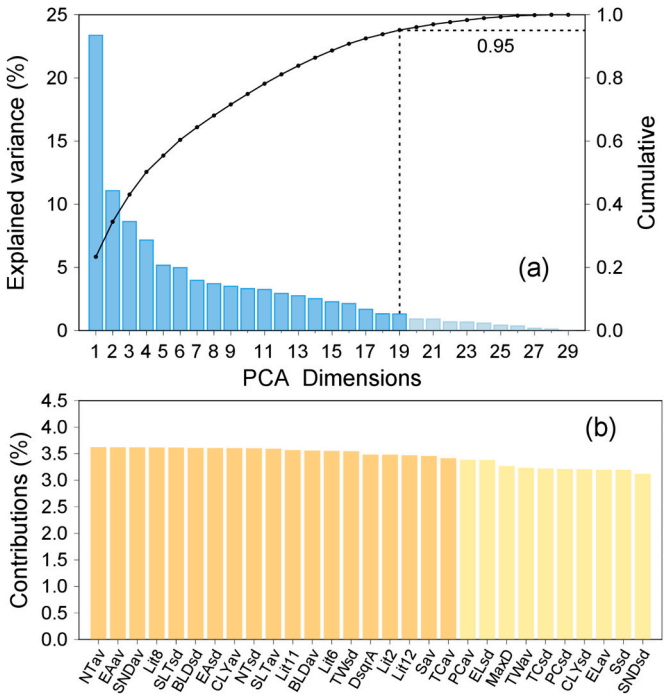


Fig. 10. (a)-(b) Result of principal component analysis (PCA) on the full dataset: (a) scree plot, *i.e.*, contributions to the total variance explained by each principal component (PC); (b) contribution of each variable to the PC decomposition in (a).

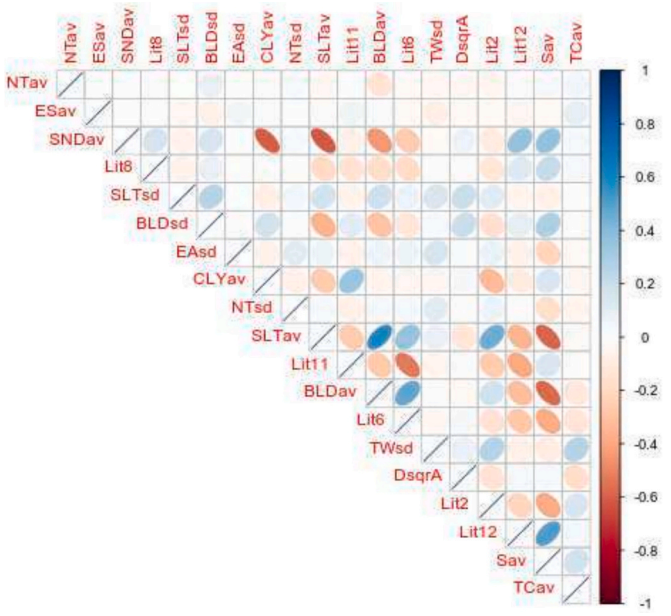


Fig. 11. Linear correlation plot of the selected 19 variables in the final benchmark dataset. Ordering is according to the decreasing contribution from principal component analysis (PCA), as in Fig. 10(b). Symbols with larger ellipticity correspond to a higher degree of correlation between the two considered variables. One can observe that, after variable reduction, most linear correlations were removed.

would like to obtain a benchmark dataset free from major correlations. That is because the classification performance of a few methods (for example LR) would be penalized by correlations, while the performance of most pure machine learning models would not be affected. Moreover, we want to reduce the chance of overfitting the data and reduce the overall dimensionality of the problem.

The process of defining and removing correlations should not be linked to a specific classification method, so that we would end up with data that can be equally useful for a fair comparison of a range of different landslide susceptibility models. It should not be linked to the values of the target variable either because we have two target variables (`p1`, `p2`) and this would give performance advantages to some methods.

The final dataset was prepared to minimize correlations between variables and to mitigate any bias in the results of part two of this experiment. We first excluded depth to bedrock and SU area in the final dataset. The depth to bedrock variable was hardly changing across the study area, due to a low number of data points used to build the (global) SoilGrids model. Slope unit area had a peculiar behavior in at least two of the independent studies conducted in step one by G7 and G9.

After removing these variables, we checked for correlations between each pair of quantities in the dataset, using the standard `corrplot` R package (Wei and Simko, 2021). A graphical representation of the test is in Fig. 9, and the figure clearly shows a large degree of correlations. To visually explore correlations beyond linear, we used the `GGally` R package (Schloerke et al., 2022) to plot two predictors against each other, for all possible pairs, and check the general trend of data with respect to the binary variables `p1` and `p2`. As the full dataset is rather large, to visualize the results we split the data into five pieces and prepared figures for each pair combination, for a total of 10 figures for each presence variable. As this is only a visual inspection of data, we presented them in the supplementary material (Figs. S1–S10). It is clear from these figures that most variable pairs have a complex relationship with one another. Moreover, it is hard at this stage to establish which variable has a distinct behavior with respect to the absence or presence of landslides, despite a few differences described by the diagonal plots.

The lithology variables clearly stand out with respect to the others, which is expected, as they are different in nature from the other variables.

We reduced the correlation between variables using information gained from principal component analysis (PCA). We chose this method because it is not biased towards any modeling method and can measure overall correlations rather than just pairwise linear correlations. Principal component analysis is an orthogonal linear transformation of data to new coordinates that concentrate the largest variance in a smaller number of axes with respect to the original coordinates (Joliffe, 2002).

Table 3
Attribute table of the final dataset. After principal component analysis (PCA) analysis, we retained the 19 variables contributing to 95 % of the total variance in the full dataset. In the table, SD stands for standard deviation and SU stands for slope unit.

Column name	Variable	Short name
`id`	Unique slope unit identifier	–
`slope_aver`	Mean Slope Steepness [deg]	Sav
`tcurv_aver`	Mean Profile Curvature	PCav
`nthns_aver`	Mean Northerness	NTav
`nthns_stdd`	SD of Northerness	NTsd
`easns_aver`	Mean Easterness	EAav
`easns_stdd`	SD of Easterness	EAsd
`twi_stddev`	SD of Topographic Wetness Index	TWsd
`BLDFIE_ave`	Mean Bulk density [kg/m^3]	BLDav
`BLDFIE_std`	SD of Bulk density [kg/m^3]	BLDsd
`CLYPPT_ave`	Mean Weight % of clay particles	CLYav
`SNDPPT_ave`	Mean Weight % of sand particles	SNDav
`SLTPPT_ave`	Mean Weight % of silt particles	SLTav
`SLTPPT_std`	SD of Weight % of silt particles	SLTsd
`D_sqrt_A`	Maximum Distance/$\sqrt{SU Area}$	MaxD
`Litho2`	Alluvial deposits (%)	Al / Lit2
`Litho6`	Unconsolidated sedimentary rocks	Usr / Lit6
`Litho8`	Marlstone	M / Lit8
`Litho11`	Schistose metamorphic rocks	Ssr / Lit11
`Litho12`	Carbonate rocks	Cr / Lit12
`presence1`	Binary landslide presence flag	`p1`
`presence2`	Binary landslide presence flag	`p2`

At variance with standard applications of PCA, we eventually did not use the transformed orthogonal variables singled out by PCA, but only the reduced set of original variables with maximal information content.

First, we performed PCA on the whole set of data, amounting to 29 variables (refer to Table 1). We used the R package `factoextra` (Kassambara and Mundt, 2020) to visualize results as in Fig. 10.

Data were normalized (R function `scale`) before feeding them to `pca()`. The percent contributions to principal components (PCs) calculated with the full dataset (Fig. 10(a)) is useful for determining how many PCs must be retained to explain a given fraction of the total variance in the data. We opted for a threshold of 95 %; accumulating the contributions, we conclude that the first 19 PCs are enough for the purpose.

Second, we considered the contributions of all original variables to the first 19 PCs (Fig. 10(b)). Explaining the required 95 % total variance would require at least 19 PCs, and we would need a minimum of 19 variables to do so with a linear combination, as in PCA. Thus, we decided to select the set of 19 variables that contributed to the 19 most relevant PCs, going from left to right in Fig. 10(b).

We validated the results of the PCA delimited dataset measuring the pairwise linear correlations among the reduced set of 19 variables. Fig. 11 shows that few correlations are left. Table 3 lists the final set of variables; note that all of the five lithological variables were retained by the PCA–based procedure.

6. Results with final dataset and constrained workflow

Results for the final dataset (Section 5, Table 3) present LSMs obtained with 16 different models. They correspond to one map for each of the 11 groups participating in the benchmark for each scenario `p1` and `p2`, except for G1 (Section 4.1), who contributed with four results as in the first step (LDA, QDA, LRM, and CFM), G7 (Section 4.7) with two results (LR and ANN_1), and G8 (Section 4.8) with two results (XGB_5 and BLR). All the remaining groups either presented one result from the very beginning, or decided to show only their best result. G11 (Section 4.11) presented results for two different methods for `p1` (STK) and `p2` (RF_3).

To compare the 16 susceptibility values, we calculated the sum of squared differences as in Eq. (2) (Fig. 12), the pairwise Pearson correlation coefficient (Fig. 13), AUC_{ROC} (Figs. 14 and 15, for `p1` and `p2`, respectively) and Brier score as in Eq. (1), (Fig. 16).

In general, metrics are much more similar to each other than in the first step; this is largely expected, given the prescription for training/CV and use of same data. Figs. 12 and 13 indicate similar considerations. A few methods stand out as more dissimilar from the others, for example QDA, RF_2 and STK (`p1`) and QDA, RF_1, and RF_2 (`p2`), in both figures. The correlation plots also show a peculiar pattern within the block of the first four results (all by G1), particularly for the combined model, CFM (which is expected.

The difference between the participant reported AUC_{ROC} values and values of AUC_{ROC} measured by the organizers are less variable compared to step one of this experiment (Figs. 14 and 15). All of the AUC_{ROC} values are between 0.7 and 0.75, are slightly higher for `p2`, and highest for the

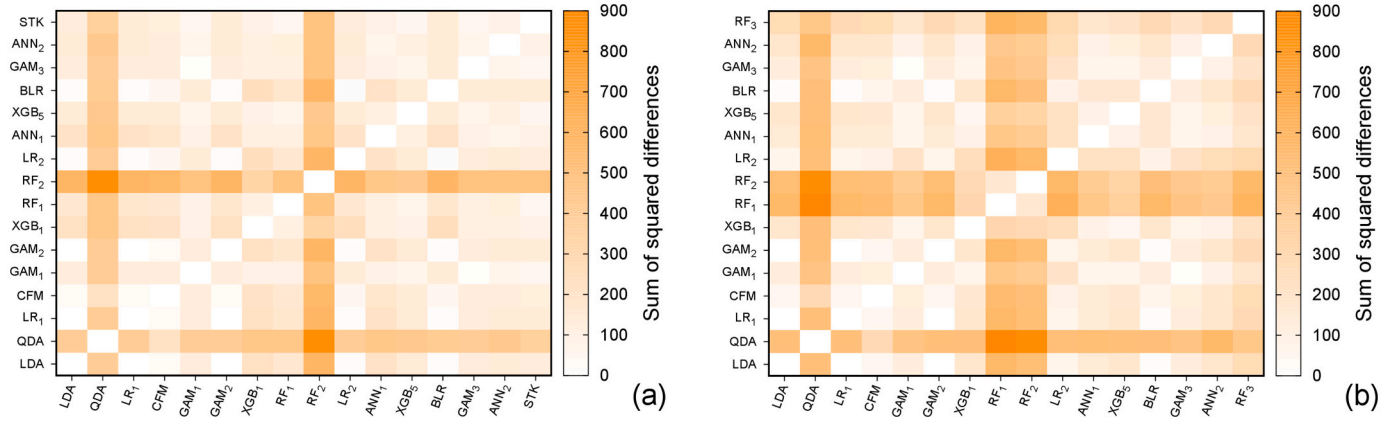


Fig. 12. Pairwise comparison of the results of different methods applied with same workflow, during step two of the experiment, described in Section 5. Numerical values calculated as in Eq. (2). Panels (a) and (b) correspond to the target variables `p1` and `p2`, respectively. Names of the different methods are as in Fig. 3, Sections 4.1–4.11.

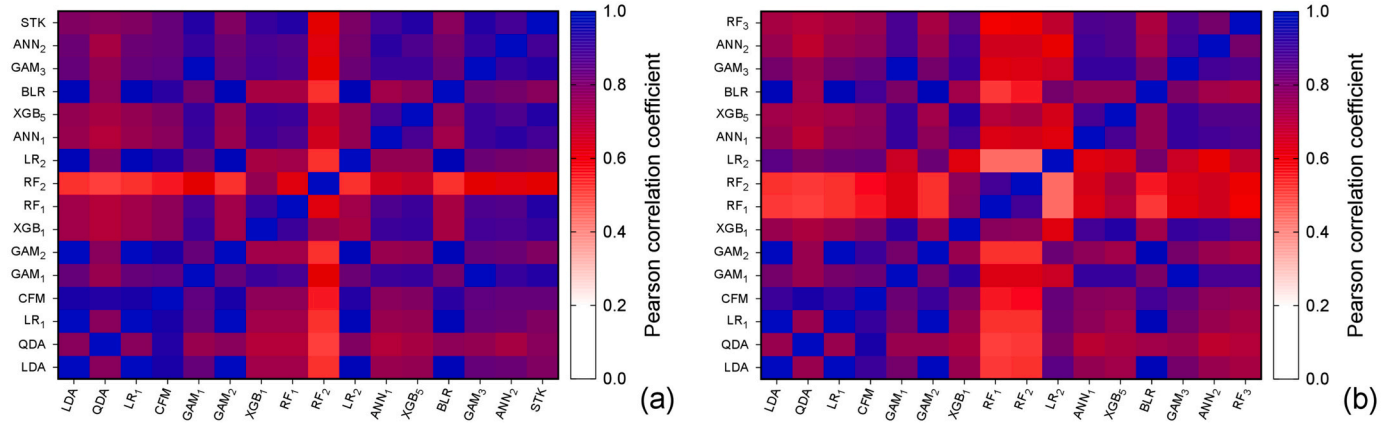


Fig. 13. As in Fig. 12, but for pairwise Pearson correlations.

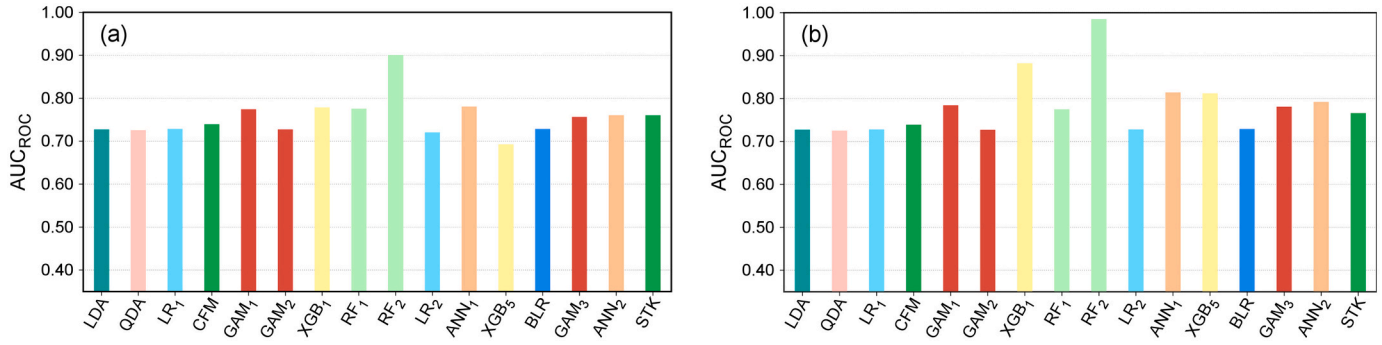


Fig. 14. Comparison of AUC_{ROC} values for the different methods considered in step two of the experiment (Section 5), limited to the `p1` landslide flag. Values in (a) were calculated by the contributors, and values in (b) were obtained by the organizers of the experiment calculating AUC_{ROC} from the attribute tables provided by the contributors. Names of the different methods are as in Fig. 3, Sections 4.1–4.11.

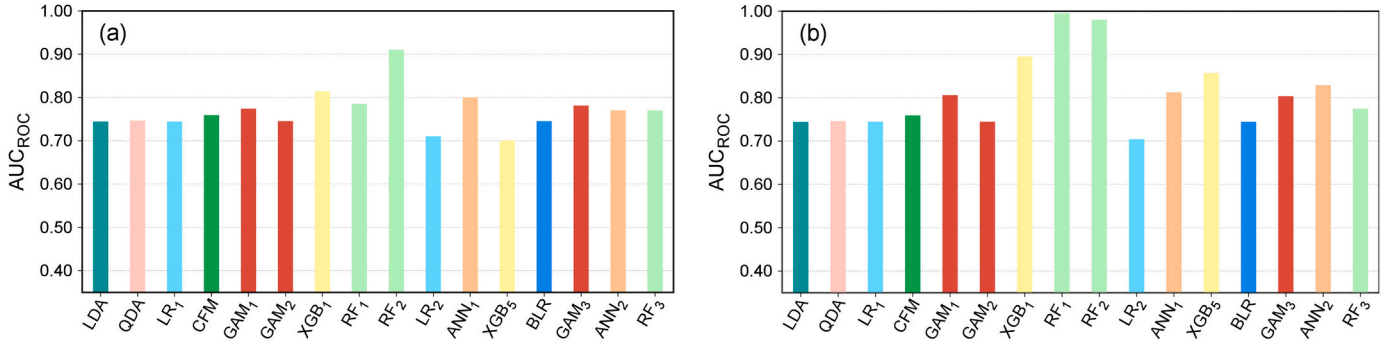


Fig. 15. As in Fig. 14, but for the `p2` landslide flag.

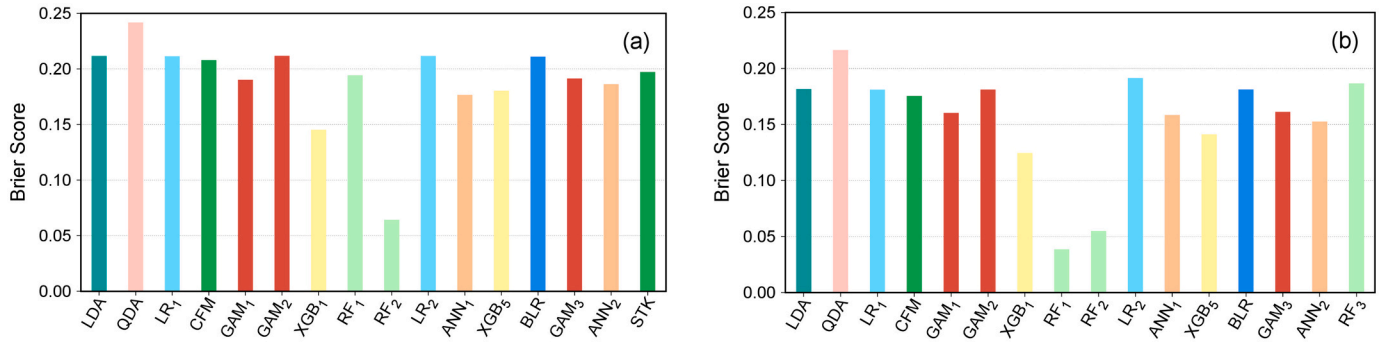


Fig. 16. Brier scores (*cf.* Eq. (1) (Brier, 1950)) for the results with the final benchmark dataset (including 10–fold CV), calculated by the organizers of the experiment. (a) is for `p1`, (b) for `p2`. Names of the methods as in Fig. 3, Sections 4.1–4.11.

RF_2, XGB_1 and RF_1 models. On average, AUC_{ROC} values are slightly lower than in step one of the experiment.

In contrast, the Brier score results measured by the organizers (Fig. 16) show larger variability between groups. The results for RF_2 still are the best ones (in this case, smaller values correspond to better agreement), and in the `p2` case the RF_1 value is practically the same. We note that the GAM values are slightly different from each other, despite using the very same training/CV strategy; the GAM_2 results are different from the other two.

We show the final landslide susceptibility maps to qualitatively investigate spatial patterns (Figs. 17 and 18, for `p1` and `p2`). Model results are float values in the [0,1] interval; in the figures we have classified each result in five equal intervals; this is one of the possible classification procedures, apt to allowing direct comparison of the maps. Boxplots of the distributions from the maps' numerical values (Fig. 19) show the clear difference between `p1`, in (a), and `p2`, in (b), as the mean values are substantially smaller for the latter (around 0.3–0.35 instead of about 0.5), but variations are larger in the second case. In both cases, a few methods show wider distributions than the others, for example the QDA, RF_1, and RF_2 stand out in this respect. Looking at same methods applied by different groups, LR seems to give consistently similar distributions, as well as the GAM (the whisker for GAM_2 looks slightly narrower) and the XGB. The RF results instead seems to give more variable distributions in the different applications.

Figs. 20 and 21 show different options for class breaks. We show, in addition to the class breaks used in the maps of Figs. 17 and 18 in panel (a), alternative classifications for all of the results, namely Jenks natural breaks in (b), and 20–40–60–80 percentiles, in (c). These two methods select breaks based on the distributions, so that break values are different for each model result. We show maps classified with Jenks breaks in the supplementary material, Figs. S11 and S12.

We acknowledge that classification into discrete categories would deserve a chapter on their own, but we deemed this beyond the aim of this review, and we preferred to focus on the distribution of values. The

values of class breaks give an alternative description of the susceptibility distributions, with respect to both maps (Figs. 17 and 18) and to boxplots (Fig. 19).

7. Discussion

We presented the first cross–examination of susceptibility modeling approaches with truly independent tests using a benchmark dataset. We surveyed interested contributors in the landslide science community, asking to prepare susceptibility maps with their methods of choice for a given dataset. The experiment was designed to be a rather specific one but, aside from quantitative calculations, we obtained a range of subjective responses on how to approach the problem, and the proposed experiment resulted in multiple outcomes.

Besides the use of different classification methods, the research groups participating in the experiment tried to answer alternative questions in addition to the simple, technical ones asked in the original survey. This triggered a second iteration in response to findings from the initial experiment, during which the dataset underwent modifications and a specific workflow was singled out.

We discuss separately the results corresponding to the preliminary dataset, presented in Sections 4.1–4.11, and the final version of the dataset, *i.e.*, our proposal for a benchmark dataset for landslide susceptibility assessment, presented in Section 5, with results in Section 6.

7.1. Discussion: preliminary dataset and results

The first part of the experiment, in which the contributors could play freely with methods and spatial variables, showed how the same algorithm could return quite different outputs, due to different model implementation techniques that mostly went unreported. This highlights the need to understand how these techniques influence model outcomes.

Variable selection was performed by 8 out of 11 participant groups, with several different methodologies, and the number of considered variables ranged from one (SU area) to all. We can distinguish different methods in two classes: (1) in a few cases groups would only use data to perform the selection, (2) in other cases, groups would also require performance assessment (LOO, VIF, other statistical tests), making the selection specific to the modeling approach. Selection based solely on data always consisted of removing collinear variables, although with different approaches for identifying them, mostly pairwise. These considerations helped in shaping the final dataset of the benchmark experiment, as described in Section 5.

One notable point was the relevance attributed to the SU area variable. In fact, in the original experiment, `area` was not even intended to be considered as a predictor. Nevertheless, G9 explicitly investigated the effect of using SU area alone, or in combination with other predictors, *versus* the case in which it was excluded. Slope unit area was a meaningful predictor for landslide occurrence. However, this correlation was assessed as a random effect rather than a causal relation. Group 7 discussed instead a related issue, namely the aggregation of variables over SU polygons (*i.e.*, zonal statistics: the process of calculating mean, standard deviation, and percentages) and the possible confounding effect of different SU area. Group 2 kept SU area in the analysis, treating it as any other covariate. The results of the performed k–fold CV showed a wide dispersion of the mean decrease in deviance explained (*i.e.*, variable importance) for both `p1` and `p2` (similar interquartile range from 2.7 % to 7.2 %), indicating a random effect of the variable on the model output.

At the data level, it is tricky to differentiate the SU area effect from potential causal contributions of other covariates because the SU area is inherently present in all covariates. Even after explicitly eliminating the `area` variable, SU area still controls the distribution of other covariates due to aggregation within SU polygons. This issue of aggregating covariates and target variables across non–uniform aerial units gives rise to the modifiable aerial unit problem, known as MAUP (Openshaw, 1984), mentioned in Alvioli et al. (2020).

We suggest that accounting for SU area effects could enhance the practical applicability and interpretation of the LSM. This is also relevant when assessing performance. We explore this point more fully later in this section. However, a model structure where landslide occurrence is represented as presence/absence does not seem ideal to visualize this relationship, and an investigation of the use of this variable in models targeting landslide counts *versus* landslide density is beyond the scope of this work and could be considered in a separate study.

Two participant groups considered the relevance of the geographical extent and location of the benchmark dataset. In fact, the dataset proposed here is an excerpt from a larger dataset used in a previous study at national level (Italy). The larger dataset, in turn, included landslide information from a national inventory. This triggered two different approaches from G2, which independently built a similar dataset for a different region in Northern Italy, and from G3, who compared the results for the benchmark subset with those at the national scale.

In the first case (G2), the average AUC_{ROC} values in the Italian Western Alps were higher than in Central Italy (up to 0.08 AUC_{ROC} points), but the average loss of performance between the training and test phases were lower in Central Italy than in the Alps (0.01 and 0.06 AUC_{ROC} points, respectively). Moreover, a different number of predictors was kept as significant in the two areas (26 in Central Italy, and only 9 in the Italian Western Alps). This indicates the need to develop more consistent models to account for spatial CV variability and to develop the model for the unique attributes within the study area.

The varying degree of completeness of the national inventory (Trigila et al., 2010) is discussed in great detail by Loche et al. (2022). This emphasizes that multiple benchmark datasets based on international open–source data are useful not only to investigate the comparative performance of mapping methods but also to explore and discuss geographical differences. For example, the landslides within this part of central Italy are dominated by translational slides. Consequently, the predictors included in this benchmark dataset were chosen to describe the attributes of the terrain that influence these landslide types within this region of Italy. Other variables can affect translational landslides, *e.g.*, terrain attitude or rock structure, which were not available for this study. Other geographic areas or other landslide types may require a different combination of predictors for a meaningful susceptibility model comparison to be conducted.

In the second case, G3 compared the national susceptibility maps (Loche et al., 2022) with the results of the benchmark dataset, and evaluated the response of fixed and random effects, as they were modeled in the same way in the two studies. They found that the non–linear variables (`slope_aver`, `Max_Distan` and `D_sqrt_A`) behave in the very same way in the two cases. Moreover, they calculated the Pearson correlation coefficient between both presence scenarios and all of Italy, which resulted in values of 0.81 and 0.83, for `p1` and `p2`, respectively. This confirms the relationship between the susceptibility zonations applied across different scales, from our small study area for the benchmark dataset up to the entire nation of Italy.

G4 and 11 explored the effects of different CV methods. Group 4 implemented several strategies, including non–spatial CV, spatial block

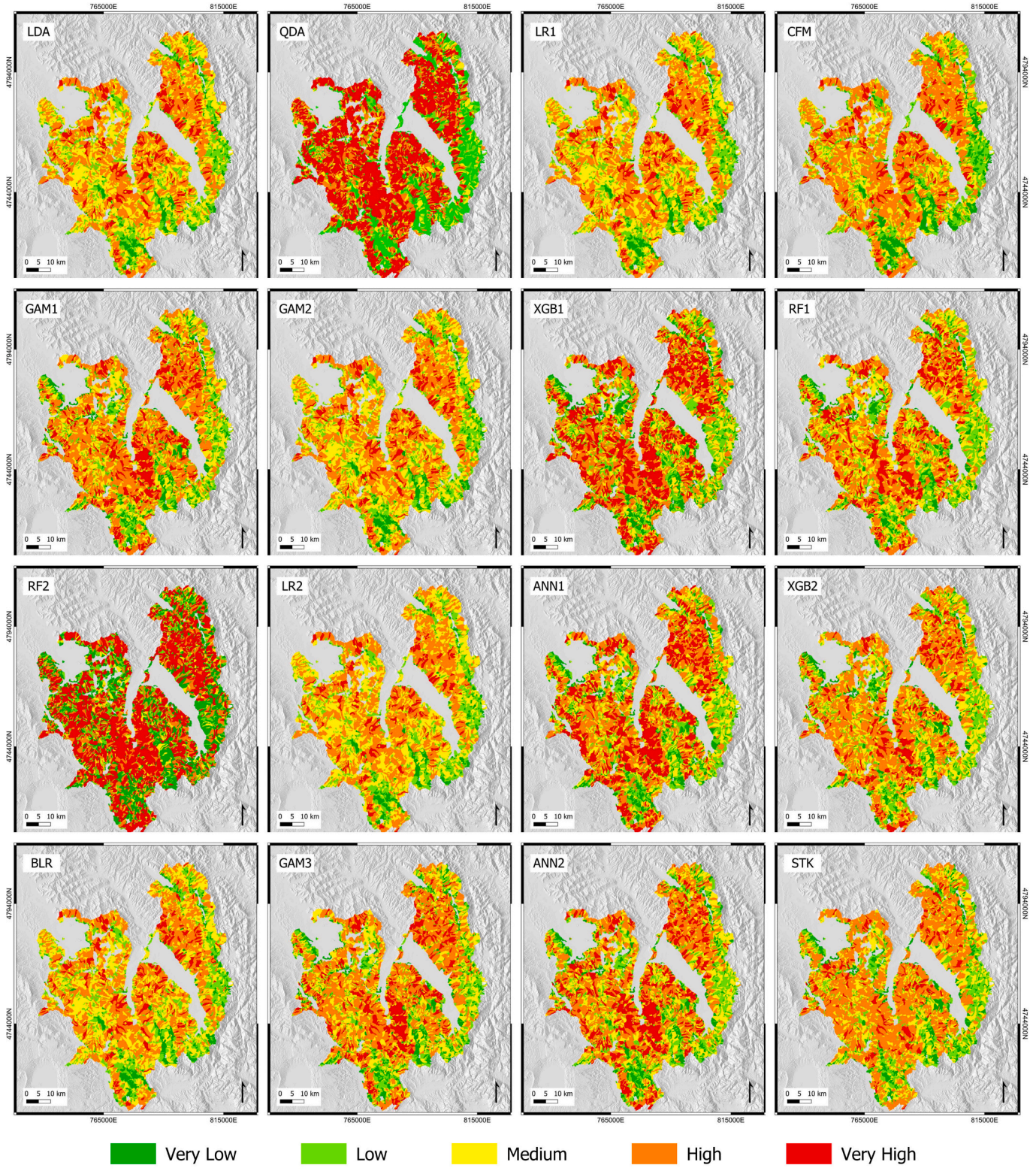


Fig. 17. Susceptibility maps corresponding to the case `p1`; model names as in Fig. 3 and Table 2. Shaded relief map as in Fig. 2.

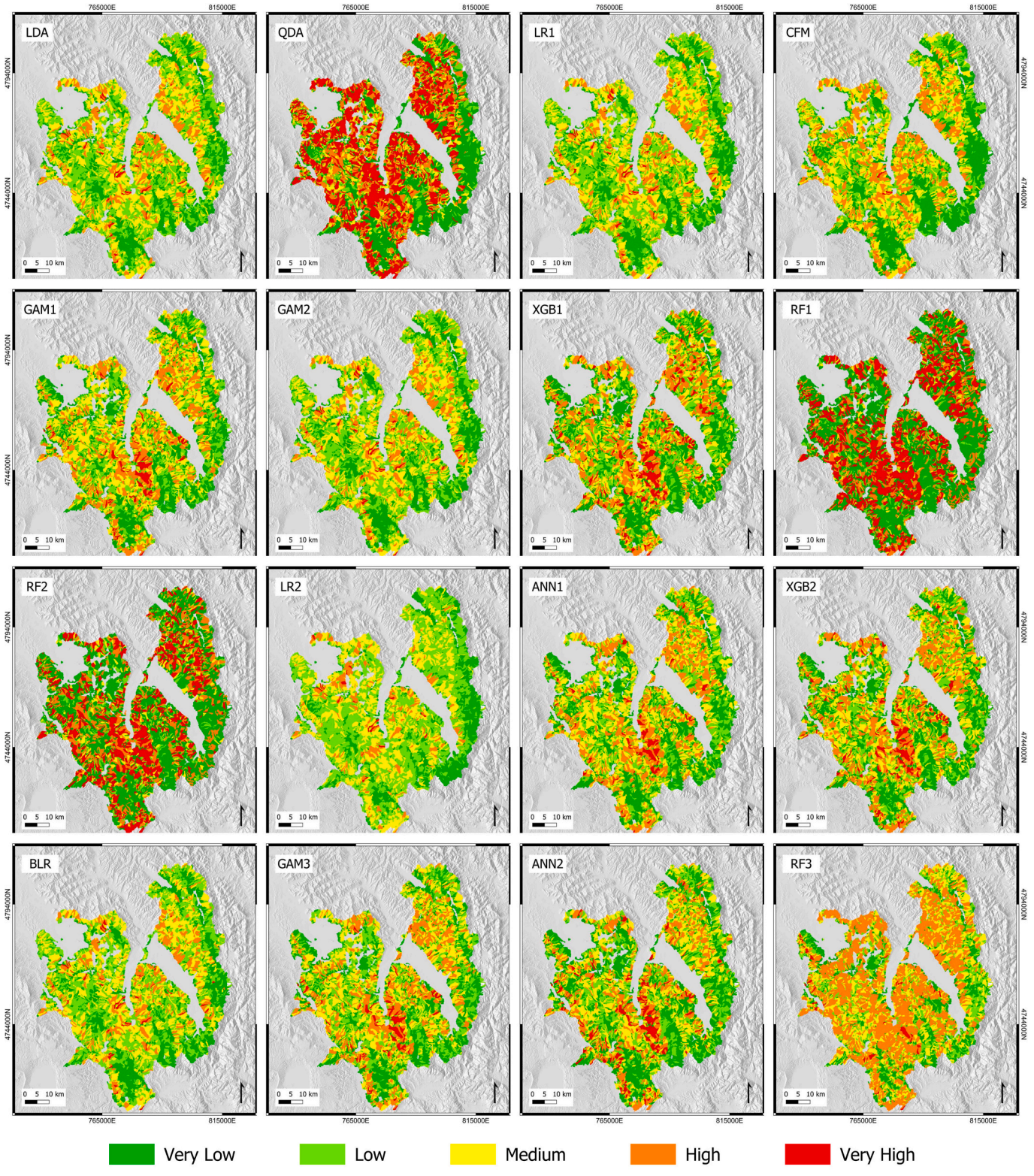


Fig. 18. Susceptibility maps corresponding to the case p2; model names as in Fig. 3 and Table 2. Shaded relief map as in Fig. 2.

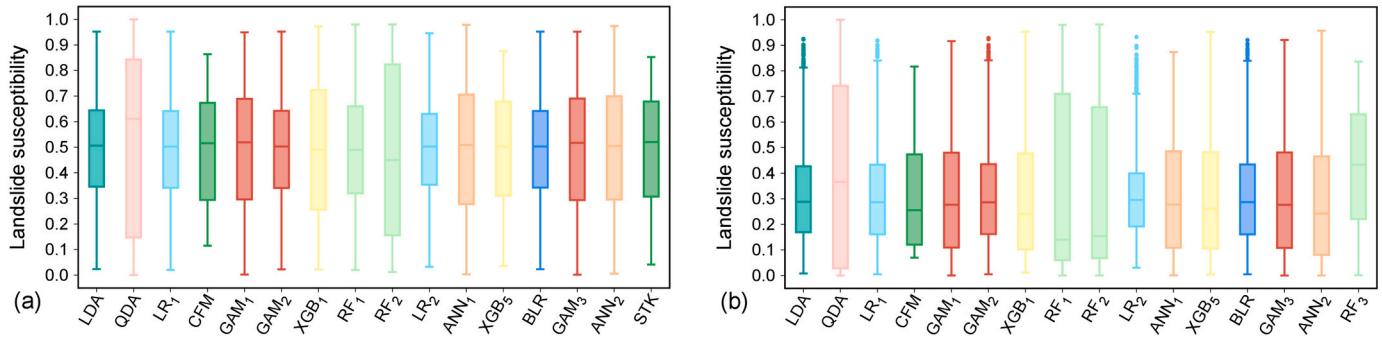


Fig. 19. Boxplots of the distributions of susceptibility values in each model; (a) for `p1`, (b) for `p2`. The box is around the region between the 1st and 3rd quartiles, horizontal line is at the median value, whiskers extend to 1.5 times the interquartile range, and the points are outliers.

CV, and spatial clustering CV, without CV based on XGBoost model. The reason to implement a CV, and a spatial CV, is to reduce overfitting, considering that spatial data is a special case of application of machine learning and conventional (random) CV does not account for spatial patterns. In general, spatial CV prioritizes consistency over accuracy.

Results (for the `p1` scenario, Fig. 6) show that the difference in AUC_{ROC} values between training and testing of the XGBoost model without CV (XGB_1) is slightly higher than that of non–spatial CV (XGB_2), spatial block CV (XGB_3), and spatial clustering CV (XGB_4). The difference in AUC_{ROC} values between training and testing of XGBoost models using non–spatial CV and spatial CV shows the same result. The XGB_6 model from G11 obtained similar testing performance with their spatial block CV scheme. In step two, G4 only provided the non–spatial CV results, as required, for the results to be exactly comparable with others.

We provide one last comment about performance assessment in the first part of the experiment. The authors of G7 stressed the need of devising a classification performance metric that would incorporate the difference in size of SUs, which is not included in any of the metrics used in this experiment. Their proposed solution focuses on using success rate, which accounts for the heterogeneous spatial extent of SUs by depicting the total area along the *x*–axis. This stands in contrast to the classical ROC curve approach, where the *x*–axis represents the false–-positive rate and treats SUs as equivalent entities. This often results in moderate to good AUC_{ROC} values, although success rate values may indicate non–informative models in the spatial context, because SU ranking based on the likelihood of landslide presence may appear spatially random.

Although we maintain that the comments about taking SU size into consideration for performance assessment are valuable, we decided to present two simple metrics in the second part of the experiment, *i.e.*, AUC_{ROC} and the Brier score. These metrics facilitate the comparison between the first and second parts of our experiment.

7.2. Discussion: final benchmark dataset and workflow

In the second part of this study, we proposed a revised dataset for LSM benchmarking and a well–defined prescription for the application of the different methods. The latter was a mandatory task because part one showed that the participants workflows is as important as the dataset itself.

In summary, there were four differences between step 1 and step 2 of this study.

- Updated dataset. We removed several input variables based on the variance calculated with a PCA analysis, and added lithological information, as suggested by several participants.
- Use of all proposed variables. No model dependence should be introduced by further variable selection based on specific model performances.
- Application of 10–fold CV without repetition, selected at random. In addition to standardizing the CV procedure, this provides LSMred s in which all of the SU susceptibility values are predicted, and not fit.
- Performance assessment with AUC_{ROC} by the contributors. In addition, an independent calculation of AUC_{ROC} and Brier score values was performed by the organizers of the experiment.

The PCA procedure used to remove correlations between the variables proved effective. We stress that in the second part of the experiment all contributors were asked to apply their methods to the original (reduced set) of variables, and not to transformed variables from PCA. This certainly reduced the possible bias introduced by using linear combinations.

Principal component analysis is an effective linear dimensionality reduction technique for capturing linear relationships, but it may not adequately capture complex, non–linear patterns inherent in the data. Conversely, one of the primary advantages of using machine learning algorithms is their capability to discern intricate relationships between features and the target variable, thereby capturing hidden non–linear patterns. This could have contributed to the overall decrease in model performance observed in step two.

The analyses indicate that the common use of CV without repetition (*e.g.*, 10–fold) can be problematic when it comes to interpreting and comparing susceptibility maps. This is because susceptibility values obtained by fitting the data are removed from the LSM, and only predicted values are retained, but in the literature those fitting data are used in different proportions.

The experiments present some notable differences between the results produced using the `p1` and `p2` presence indicators. Generally, the `p2` indicator results in better AUC_{ROC} (Figs. 14 and 15) and Brier scores (Fig. 16) compared to the `p1` indicator. However, the `p2` indicator results also show more variance (Fig. 22) and lower probabilities (Fig. 18). These trends are likely due to the `p2` indicator allowing SUs with one landslide point to be considered as not containing a landslide. The improved model metrics for the `p2` indicator may be due to the reduced number of SUs with landslides, which can artificially improve our chosen metrics (Davis and Goadrich, 2006). Allowing a single landslide

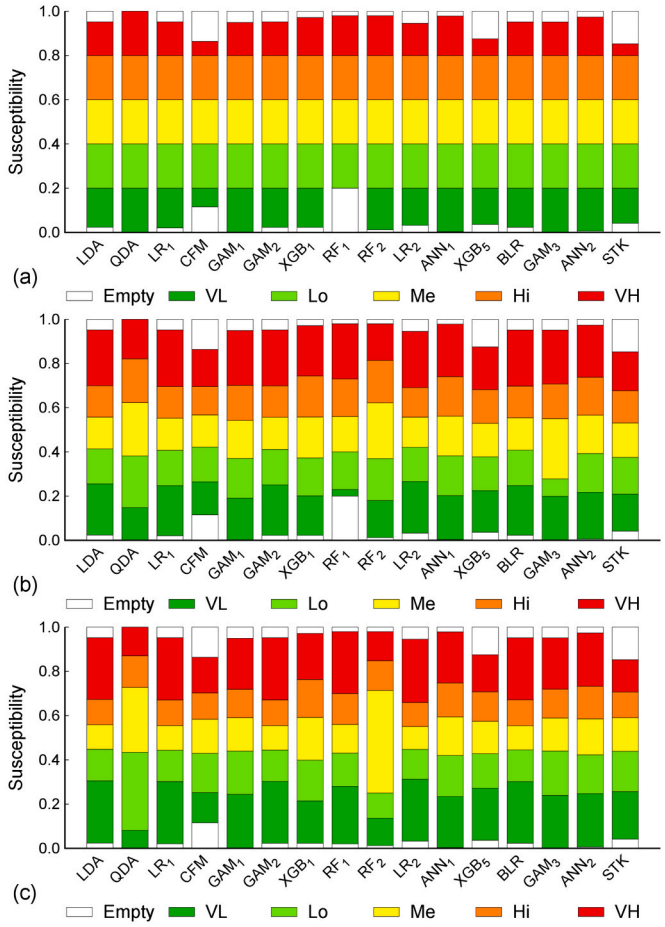


Fig. 20. Stacked bars representing different class breaks, for the `p1` scenario, using (a) equal intervals (same breaks, different number of values in each class; used in Fig. 17), (b) Jenks natural breaks (different breaks and different number of values in each class), and (c) percentiles of the distributions (different breaks, but each class contains the same number of values, *i.e.*, of slope units). In all cases, we show the actual lower and upper limits in the distribution of susceptibility values with a white band.

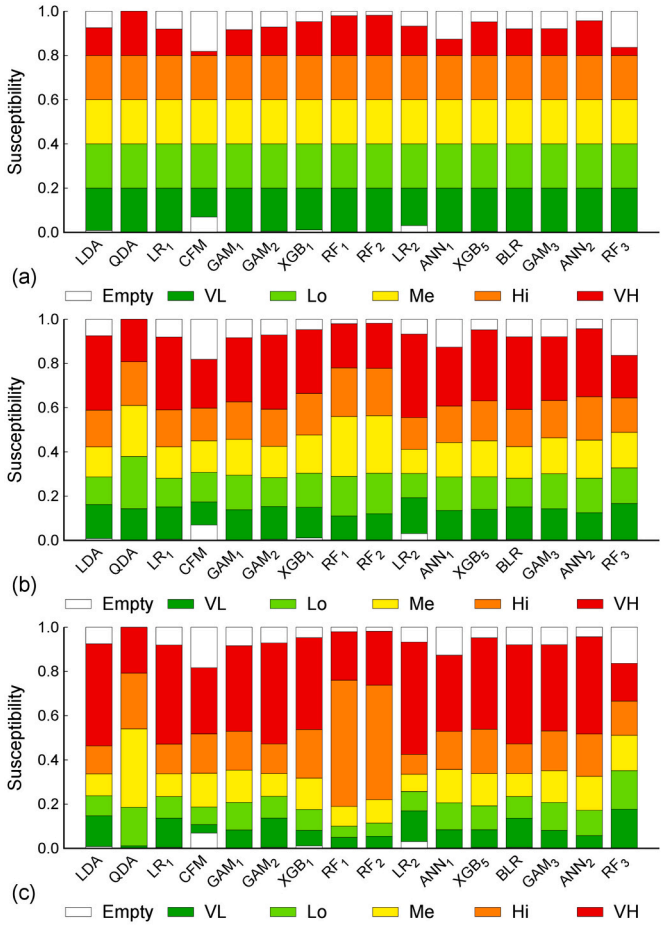


Fig. 21. As in Fig. 20, but for the `p2` scenario; breaks in (a) correspond to the classification of Fig. 18.

point to not change the presence indicator in `p2` may also increase the variance in the model results. This is because the attributes of SUs containing one or many landslides may be similar, hampering a model's ability to differentiate between landslide and non–landslide SUs. The `p1` indicator avoids this issue by not allowing any landslide points within the SUs labeled as a non–landslide slope unit. The reduced area indicating landslide presence (Fig. 2) also explains the reduction in probabilities using the `p2` flag. In summary, these results do not favor one method of categorizing SUs as containing a landslide. However, the chosen method may have notable effects on the model results.

Finally, we stress that we did not intend to select the "best" method here, as our only aim is to establish a benchmark dataset and workflow, that could be useful as a standard reference for calculations by other scholars. Different methods may be more useful in diverse settings, and different predictors may be available in other areas, with respect to those considered here. We note that the standardized workflow leveled out model performances, as Figs. 14–16 show, with respect to Figs. 6 and 7. Nevertheless, the spatial patterns in Figs. 17 and 18 show clear differences that one could not grasp from the numerical values of the performance metrics. The residual differences can be ascribed to different predictive abilities of the models in the study area presented here as a benchmark. Fig. 22 shows the variance of results across the 16 different methods, in each SU, colorized in five classes. We observe that variability is larger for the second scenario, and has a different spatial pattern, which highlights the relevance of the method adopted to quantify landslide presence, starting from point landslide locations.

The pronounced leveling with the final dataset combined with the observation that results are most similar within each contributing group highlights that user–caused variability in model performance and quality is significant. Different software packages, coding styles, user error, and unclear workflows can all lead to significant model differences that are difficult to parse out, as evidenced by step one of this experiment. This further highlights the benefit of a benchmark dataset for directly comparing future susceptibility modeling approaches and emphasizes that direct comparison between models produced by different researchers without a standardized workflow should be done with caution.

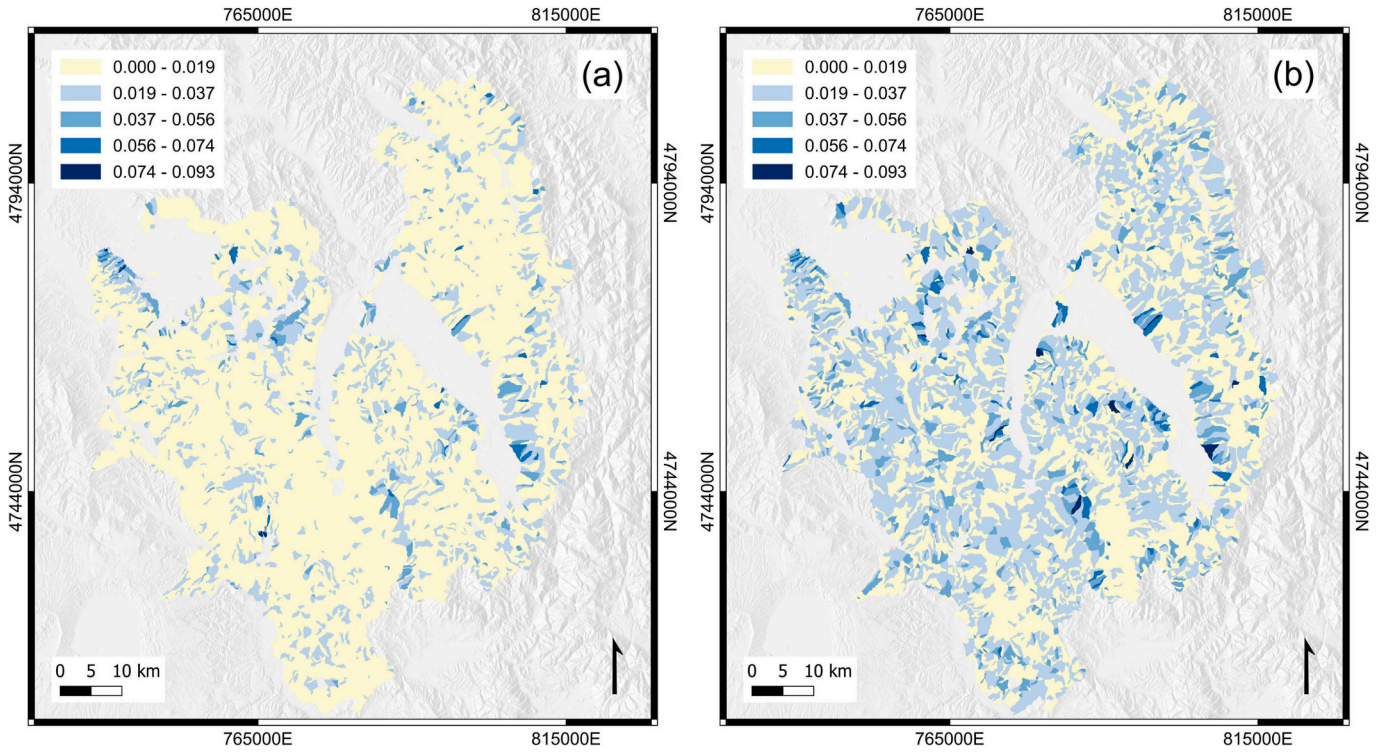


Fig. 22. Variance of the susceptibility values in each slope unit, obtained with the benchmark dataset proposed here for 16 different classification methods. (a) Variance of results shown in Fig. 17, for the landslide presence scenario `p1`, and (b) scenario `p2`, shown in Fig. 18. Shaded relief map as in Fig. 2.

8. Conclusions

We presented the first participatory experiment to systematically investigate current methods and practices in landslide susceptibility modeling. Our results clearly highlight the benefit for landslide scientists and practitioners to benchmark their results against known data, methods, and workflows.

We demonstrated that two fundamental steps for obtaining robust and reproducible LSMs are (1) a critical selection of model input and detailed rules for application of a classification method (referred to as "workflow" throughout this work), and (2) a critical evaluation of the output. Our findings also highlight the benefit not only of benchmark datasets, but also of very specific, unambiguous, and shared guidelines on how to build an LSM. These warrant being included in guidance ranging from data collection, data selection and pre–processing, to selection and application of methods and, eventually, to the assessment of results with proper metrics.

We designed the dataset to serve exactly this purpose, and the results of the experiment demonstrate the success of our objectives. Moreover, similar datasets and results for different regions of the world, specific to landslide science, would improve our understanding of landslide susceptibility predictions, and would enable their real–world applicability.

Data and code availability

The final benchmark dataset corresponds to the same SUs set as in the preliminary dataset, with a modified attribute table. In addition, we provide in separate vector layers the results for all of the methods used in step two of the experiment, used to prepare Figs. 11–21. We also share code to (1) calculate susceptibility value within the GAM model of Section 4.3, and (2) calculate AUC_{ROC} values and Brier scores from the attribute table of the vector layer with results.

The benchmark dataset, results, and code are publicly available for download at the main CNR IRPI SU project page, at: `https://geomorphology.irpi.cnr.it/tools/slope-units`, under the section `Data → Benchmark Dataset`. We provide the datasets in OGC GeoPackage vector format (GeoPackage is an open, standards-based, platform–independent, portable, self–describing, compact format for transferring geospatial information) and in ESRI Shapefile format. The maps are identical, both of them have the same attributes, and are in EPSG:32632 - WGS 84 / UTM zone 32 N projected reference system. Code is a combination of bash scripting, GRASS GIS, and R.

Authors contributions

M. Alvioli: designed the experiment, prepared the dataset (DD), collected results from participants, wrote the first draft (WD), revised the manuscript (RM). M. Loche: DD, contributed as G3, and RM. L. Jacobs: WD and RM. C. H. Grohmann: RM. M. T. Abraham: G10 and RM. K. Gupta and N. Satyam: G10. G. Scaringi: G3. T. Bornaetxea and M. Rossi: G1. I. Marchesini: G3 and RM. L. Lombardo: G3. M. Moreno: G9 and RM. S. Steger: G9. C. Camera: G2 and RM. G. Bajni: G2. G. Samodra, E. E. Wahyudi and N. Susyanto: G4 and RM. M. Sinčić: G5 and RM. S. Bernat Gazibara: G5. F. Sirbu: G6 and RM. J. Torizin and N. Schüßler: G7 and RM. B. Mirus and J. Woodard: G8, RM, proof–read English and took care of internal USGS revision. Héctor Aguilera Alonso: G11 and RM. J. S. Rivera–Rivera: G11.

Disclaimer

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to my data in the manuscript. Will attach files here, if possible.

Acknowledgments

M. Alvioli partially was supported by Italian MUR under URGERE, 2023–2025, project no. B53D23007260006. C.H. Grohmann was funded by FAPESP (grant #2023/11197-1) and CNPq (grant #311209/2021-1). M. T. Abraham was supported by Alexander von Humboldt foundation with a post–doctoral fellowship. T. Bornaetxea was supported by the HGI research group of the University of the Basque Country. Marko Sinčić was supported by the Croatian Science Foundation under the project HRZZ DOK–2020–01–2432. J. Woodard and B.B. Mirus were funded by the U.S. Geological Survey Landslide Hazards Program. J. S. Rivera-Rivera was supported by the Training of Research Personnel (PRE2021-100044) with a pre–doctoral grant funded by 10.13039/501100011033, and SARAI project PID2020-116540RB-C22, funded by MCIN/AEI/10.13039/501100011033.

Appendix A. Supplementary data

Supplementary material to this article can be found online at https://doi.org/10.1016/j.earscirev.2024.104927.

References

Agterberg, F., Bonham-Carter, G., Wright, D., 1990. Statistical pattern integration for mineral exploration. In: Gaál, G., Merriam, D. (Eds.), Computer Applications in Resource Estimation, Computers and Geology. Pergamon, Amsterdam, pp. 1–21. https://doi.org/10.1016/B978-0-08-037245-7.50006-8. Geological Survey of Canada Contribution No. 24088.

Aguilera, Q., Lombardo, L., Tanyas, H., Lipani, A., 2022. On the prediction of landslide occurrences and sizes via Hierarchical Neural Networks. Stoch. Env. Res. Risk A. 36, 2031–2048. https://doi.org/10.1007/s00477-022-02215-0.

Aguilera, H., Rivera Rivera, J.S., Guardiola-Albert, C., Béjar-Pizarro, M., 2023. Ensemble learning on the benchmark dataset for landslide susceptibility zonation in Central Italy. In: EGU General Assembly 2023. https://doi.org/10.5194/egusphere-egu23-16251. Vienna, Austria, 24–28 April.

Akgun, A., 2012. A comparison of landslide susceptibility maps produced by logistic regression, multi-criteria decision, and likelihood ratio methods: a case study at İzmir, Turkey. Landslides 9, 93–106. https://doi.org/10.1007/s10346-011-0283-7.

Akgun, A., Dag, S., Bulut, F., 2008. Landslide susceptibility mapping for a landslide–prone area (Findikli, NE of Turkey) by likelihood–frequency ratio and weighted linear combination models. Environ. Geol. 54, 1127–1143. https://doi.org/10.1007/s00254-007-0882-8.

Ali, M., 2020. PyCaret: an open source, low-code machine learning library in Python. URL: https://www.pycaret.org. pyCaret version 1.0.

Allen, M.P., 1997. Understanding Regression Analysis. Springer US. https://doi.org/10.1007/b102242.

Alvioli, M., Marchesini, I., Reichenbach, P., Rossi, M., Ardizzone, F., Fiorucci, F., Guzzetti, F., 2016. Automatic delineation of geomorphological slope units with r.slopeunits v1.0 and their optimization for landslide susceptibility modeling. Geosci. Model Dev. 9, 3975–3991. https://doi.org/10.5194/gmd-9-3975-2016.

Alvioli, M., Guzzetti, F., Marchesini, I., 2020. Parameter–free delineation of slope units and terrain subdivision of Italy. Geomorphology 358, 107124. https://doi.org/10.1016/j.geomorph.2020.107124.

Amato, G., Eisank, C., Castro-Camilo, D., Lombardo, L., 2019. Accounting for covariate distributions in slope–unit–based landslide susceptibility models. A case study in the alpine environment. Eng. Geol. 260, 105237 https://doi.org/10.1016/j.enggeo.2019.105237.

Amato, G., Fiorucci, M., Martino, S., Lombardo, L., Palombi, L., 2023. Earthquake–triggered landslide susceptibility in Italy by means of artificial neural network. Bull. Eng. Geol. Environ. 82, 160. https://doi.org/10.1007/s10064-023-03163-x.

Atkinson, P.M., Massari, R., 1998. Generalised linear modelling of susceptibility to landsliding in the Central Apennines, Italy. Comput. Geosci. 24, 373–385. https://doi.org/10.1016/S0098-3004(97)00117-9.

Bajni, G., Camera, C., Brenning, A., Apuani, T., 2022. Assessing the utility of regionalized rock–mass geomechanical properties in rockfall susceptibility modelling in an alpine environment. Geomorphology 415, 108401. https://doi.org/10.1016/j.geomorph.2022.108401.

Bajni, G., Camera, C.A., Apuani, T., 2023. A novel dynamic rockfall susceptibility model including precipitation, temperature and snowmelt predictors: a case study in Aosta Valley. Landslides. https://doi.org/10.1007/s10346-023-02091-x.

Batista, G.E., Prati, R.C., Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor. Newslett. 6, 20–29. https://doi.org/10.1145/1007730.1007735.

Beigaitė, R., Mechenich, M., Žliobaitė, I., 2022. Spatial cross–validation for globally distributed data. In: Pascal, P., Ienco, D. (Eds.), Discovery Science. Springer Nature, Switzerland, Cham, pp. 127–140. https://doi.org/10.1007/978-3-031-18840-4_10.

Bivand, R., Keitt, T., Rowlingson, B., Pebesma, E., Sumner, M., Hijmans, R., Baston, D., Rouault, E., Warmerdam, F., Ooms, J., Rundel, C., 2023. rgdal: Bindings for the 'Geospatial' Data Abstraction Library. URL: https://cran.r-project.org/web/packages/rgdal (Version 1.6–7).

Bonham-Carter, G.F., Agterberg, F.P., Wright, D.F., 1990. Weights of evidence modelling: A new approach to mapping mineral potential. In: Agterberg, F.P., Bonham-Carter, G.F. (Eds.), Statistical Applications in the Earth Sciences. Geological Survey of Canada, pp. 171–183. https://doi.org/10.4095/128059, 604 p.

Bordoni, M., Galanti, Y., Bartelletti, C., Persichillo, M.G., Barsanti, M., Giannecchini, R., D'Amato Avanzi, G., Cevasco, A., Brandolini, P., Galve, J., Meisina, C., 2020. The influence of the inventory on the determination of the rainfall–induced shallow landslides susceptibility using generalized additive models. CATENA 193, 104630. https://doi.org/10.1016/j.catena.2020.104630.

Bornaetxea, T., Rossi, M., Marchesini, I., Alvioli, M., 2018. Effective surveyed area and its role in statistical landslide susceptibility assessments. Nat. Hazards Earth Syst. Sci. 18, 2455–2469. https://doi.org/10.5194/nhess-18-2455-2018.

Bornaetxea, T., Remondo, J., Bonachea, J., Valenzuela, P., 2023a. Exploring available landslide inventories for susceptibility analysis in Gipuzkoa province (Spain). Nat. Hazards 118, 2513–2542. https://doi.org/10.1007/s11069-023-06103-w.

Bornaetxea, T., Yazdani, M., Rossi, M., 2023b. Application of the LAND–SUITE software with a benchmark dataset for landslide susceptibility zonation. In: EGU General Assembly 2023. https://doi.org/10.5194/egusphere-egu23-2283. Vienna, Austria, 24–28 April.

Brabb, E., Pampeyan, H., Bonilla, M., 1972. MG 1972. Landslide susceptibility in San Mateo County, California. U.S. Geological Survey Miscellaneous Field Studies Map MF–360, scale 1. https://doi.org/10.3133/mf360.

Bragagnolo, L., da Silva, R., Grzybowski, J., 2020. Landslide susceptibility mapping with r.landslide: a free open–source GIS–integrated tool based on artificial neural networks. Environ. Model Softw. 123, 104565 https://doi.org/10.1016/j.envsoft.2019.104565.

Breiman, L., 2011. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Brenning, A., 2008. Statistical geocomputing combining R and SAGA: The example of landslide susceptibility analysis with generalized additive models. In: Böhner, J., Blaschke, T., Montanarella, L. (Eds.), SAGA – Second out. Universität Hamburg Institut für Geographie. Volume 19 of Hamburger Beiträge zur Physischen Geographie und Landschaftsökologie, pp. 23–32, 410.

Brenning, A., 2012. Spatial cross–validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In: 2012 IEEE International Geoscience and Remote Sensing Symposium, pp. 5372–5375. https://doi.org/10.1109/IGARSS.2012.6352393.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78, 1–3. https://doi.org/10.1175/1520-0493(1950)078\%3C0001:VOFEIT\%3E2.0.CO;2.

Bucci, F., Santangelo, M., Fongo, L., Alvioli, M., Cardinali, M., Melelli, L., Marchesini, I., 2022. A new digital lithological map of Italy at 1:100,000 scale for geo–mechanical modelling. Earth Syst. Sci. Data 14, 4129–4151. https://doi.org/10.5194/essd-14-4129-2022.

Budimir, M., Atkinson, P., Lewis, H., 2015. A systematic review of landslide probability mapping using logistic regression. Landslides 12, 419–436. https://doi.org/10.1007/s10346-014-0550-5.

Buiter, S., Schreurs, G., Albertz, M., Gerya, T., Kaus, B., Landry, W., le Pourhiet, L., Mishin, Y., Egholm, D., Cooke, M., Maillot, B., Thieulot, C., Crook, T., May, D., Souloumiac, P., Beaumont, C., 2016. Benchmarking numerical models of brittle thrust wedges. J. Struct. Geol. 92, 140–177. https://doi.org/10.1016/j.jsg.2016.03.003.

Camera, C., Bajni, G., 2023. Comparison of the effectiveness of application of gams for landslide susceptibility modelling in Apennine and Alpine areas. In: EGU General Assembly 2023. https://doi.org/10.5194/egusphere-egu23-7907. Vienna, Austria, 24–28 April.

Camera, C.A., Bajni, G., Corno, I., Raffa, M., Stevenazzi, S., Apuani, T., 2021. Introducing intense rainfall and snowmelt variables to implement a process–related non–stationary shallow landslide susceptibility analysis. Sci. Total Environ. 786, 147360 https://doi.org/10.1016/j.scitotenv.2021.147360.

Carrara, A., Cardinali, M., Detti, R., Guzzetti, F., Pasqui, V., Reichenbach, P., 1991. GIS techniques and statistical models in evaluating landslide hazard. Earth Surf. Process. Landf. 16, 427–445. https://doi.org/10.1002/esp.3290160505.

Carrara, A., Cardinali, M., Guzzetti, F., Reichenbach, P., 1995. GIS technology in mapping landslide hazard. In: Carrara, A., Guzzetti, F. (Eds.), Geographical Information Systems in Assessing Natural Hazards, Volume 5 of Advances in Natural and Technological Hazards Research. Kluwer Academic Publishers, pp. 135–175. https://doi.org/10.1007/978-94-015-8404-3_8.

Chalkias, C., Polykretis, C., Karymbalis, E., Soldati, M., Ghinoi, A., Ferentinou, M., 2020. Exploring spatial non–stationarity in the relationships between landslide susceptibility and conditioning factors: a local modeling approach using geographically weighted regression. Bull. Eng. Geol. Environ. 79, 2799–2814. https://doi.org/10.1007/s10064-020-01733-x.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data

Mining. Association for Computing Machinery, New York, NY, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785.
Chen, W., Chai, H., Sun, X., Wang, Q., Ding, X., Hong, H., 2016. A GIS–based comparative study of frequency ratio, statistical index and weights–of–evidence models in landslide susceptibility mapping. Arab. J. Geosci. 9, 1–16. https://doi.org/10.1007/s12517-015-2150-7.
Chung, C.J., Fabbri, A., 1999. Probabilistic prediction models for landslide hazard mapping. Photogramm. Eng. Remote. Sens. 65, 1389–1399. URL: http://www.asprs.org/a/publications/pers/99journal/december/1999_dec_1389-1399.pdf.
Chung, C.J., Fabbri, A., 2003. Validation of spatial prediction models for landslide hazard mapping. Nat. Hazards 30, 451–472. https://doi.org/10.1023/B:NHAZ.0000007172.62651.2b.
Dahal, A., Lombardo, L., 2023. Explainable artificial intelligence in geoscience: a glimpse into the future of landslide susceptibility modeling. Comput. Geosci. 176, 105364 https://doi.org/10.1016/j.cageo.2023.105364.
Dahal, A., Tanyas, H., van Westen, C., van der Meijde, M., Mai, P.M., Huser, R., Lombardo, L., 2024. Space–time landslide hazard modeling via ensemble neural networks. Nat. Hazards Earth Syst. Sci. 24, 823–845. https://doi.org/10.31223/X5B075.
Das, S., Sarkar, S., Kanungo, D., 2022. A critical review on landslide susceptibility zonation: recent trends, techniques, and practices in Indian Himalaya. Nat. Hazards 115, 23–72. https://doi.org/10.1007/s11069-022-05554-x.
Davis, J., Goadrich, M., 2006. The relationship between precision–recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. Association for Computing Machinery, New York, NY, USA, pp. 233–240. https://doi.org/10.1145/1143844.1143874.
Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and their Application. In: Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. https://doi.org/10.1017/CBO9780511802843.
Di Napoli, M., Tanyas, H., Castro-Camilo, D., Calcaterra, D., Cevasco, A., Di Martire, D., Pepe, G., Brandolini, P., Lombardo, L., 2023. On the estimation of landslide intensity, hazard and density via data–driven models. Nat. Hazards 119, 1513–1530. https://doi.org/10.1007/s11069-023-06153-0.
Dias, H., Grohmann, C., 2024. Standards for shallow landslide identification in Brazil: Spatial trends and inventory mapping. J. S. Am. Earth Sci. 135, 104805 https://doi.org/10.1016/j.jsames.2024.104805.
Dias, H., Hölbling, D., Grohmann, C., 2021. Landslide susceptibility mapping in Brazil: a review. Geosciences 11. https://doi.org/10.3390/geosciences11100425.
Elia, L., Castellaro, S., Dahal, A., Lombardo, L., 2023. Assessing multi–hazard susceptibility to cryospheric hazards: lesson learnt from an Alaskan example. Sci. Total Environ. 898, 165289 https://doi.org/10.1016/j.scitotenv.2023.165289.
Erener, A., Sivas, A.A., Selcuk-Kestel, A.S., Düzgün, H.S., 2017. Analysis of training sample selection strategies for regression-based quantitative landslide susceptibility mapping methods. Comput. Geosci. 104, 62–74. https://doi.org/10.1016/j.cageo.2017.03.022.
Ermini, L., Catani, F., Casagli, N., 2005. Artificial neural networks applied to landslide susceptibility assessment. Geomorphology 66, 327–343. https://doi.org/10.1016/j.geomorph.2004.09.025.
Fang, Z., Wang, Y., van Westen, C., Lombardo, L., 2023. Space–time landslide susceptibility modeling based on data–driven methods. Math. Geosci. 56, 1335–1354. https://doi.org/10.1007/s11004-023-10105-6.
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.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.
Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugenics 7, 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.
GitHub, M.A., 2022. Awesome Public Datasets. URL: https://github.com/awesomedata/awesome-public-datasets.
Goetz, J., Guthrie, R., Brenning, A., 2011. Integrating physical and empirical landslide susceptibility models using generalized additive models. Geomorphology 129, 376–386. https://doi.org/10.1016/j.geomorph.2011.03.001.
Goetz, J., Brenning, A., Petschko, H., Leopold, P., 2015. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Comput. Geosci. 81, 1–11. https://doi.org/10.1016/j.cageo.2015.04.007.
Gong, G., 2006. Appendix B: Cross–Validation, the Jackknife, and the Bootstrap: Excess Error Estimation in Forward Logistic Regression. John Wiley & Sons, Ltd, pp. 205–218. https://doi.org/10.1002/0471998524.app2.
Guzzetti, F., Carrara, A., Cardinali, M., Reichenbach, P., 1999. Landslide hazard evaluation: a review of current techniques and their application in a multi–scale study, Central Italy. Geomorphology 31, 181–216. https://doi.org/10.1016/S0169-555X(99)00078-1.
Guzzetti, F., Reichenbach, P., Ardizzone, F., Cardinali, M., Galli, M., 2006. Estimating the quality of landslide susceptibility models. Geomorphology 81, 166–184. https://doi.org/10.1016/j.geomorph.2006.04.007.
Guzzetti, F., Mondini, A., Cardinali, M., Fiorucci, F., Santangelo, M., Chang, K.T., 2012. Landslide inventory maps: New tools for an old problem. Earth Sci. Rev. 112, 42–66. https://doi.org/10.1016/j.earscirev.2012.02.001.
Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with Numpy. Nature 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2.
Heckmann, T., Gegg, K., Gegg, A., Becht, M., 2014. Sample size matters: investigating the effect of sample size on a logistic regression susceptibility model for debris flows. Nat. Hazards Earth Syst. Sci. 14, 259–278. https://doi.org/10.5194/nhess-14-259-2014.
Hengl, T., de Jesus, J.M., Heuvelink, G.B., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., et al., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One 12, e0169748. https://doi.org/10.1371/journal.pone.0169748.
Ho, T., 1995. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, pp. 278–282. https://doi.org/10.1109/ICDAR.1995.598994.
Hong, H., Miao, Y., Liu, J., Zhu, A.X., 2019. Exploring the effects of the design and quantity of absence data on the performance of random forest–based landslide susceptibility mapping. Catena 176, 45–64. https://doi.org/10.1016/j.catena.2018.12.035.
Hosmer Jr., D., Lemeshow, S., Sturdivant, R., 2013. Applied logistic regression. In: Wiley Series in Probability and Statistics. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118548387
Huang, Y., Zhao, L., 2018. Review on landslide susceptibility mapping using support vector machines. Catena 165, 520–529. https://doi.org/10.1016/j.catena.2018.03.003.
Huang, F., Cao, Z., Guo, J., Jiang, S.H., Li, S., Guo, Z., 2020. Comparisons of heuristic, general statistical and machine learning models for landslide susceptibility prediction and mapping. Catena 191, 104580. https://doi.org/10.1016/j.catena.2020.104580.
Huang, F., Xiong, H., Jiang, S.H., Yao, C., Fan, X., Catani, F., Chang, Z., Zhou, X., Huang, J., Liu, K., 2024. Modelling landslide susceptibility prediction: a review and construction of semi–supervised imbalanced theory. Earth Sci. Rev. 104700 https://doi.org/10.1016/j.earscirev.2024.104700.
ISRIC, 2024. Wosis, world soil profile database. https://www.isric.org/. Accessed: July 19, 2024.
Ivakhnenko, A., Lapa, V., 1967. Cybernetics and forecasting techniques. In: McDonough, R. (Ed.), Modern Analytic and Computational Methods in Science and Mathematics, vol. 8. American Elsevier/Elsevier Publishing Co pp. xxvii + 168. Translated by Scripta Technica, Inc.
Jacobs, L., Kervyn, M., Reichenbach, P., Rossi, M., Marchesini, I., Alvioli, M., Dewitte, O., 2020. Regional susceptibility assessments with heterogeneous landslide information: Slope unit- vs. pixel-based approach. Geomorphology 356, 107084. https://doi.org/10.1016/j.geomorph.2020.107084.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning with Applications in R. Springer. https://doi.org/10.1007/978-1-4614-7138-7.
Jia, G., Alvioli, M., Gariano, S.L., Marchesini, I., Guzzetti, F., Tang, Q., 2021. A global landslide non–susceptibility map. Geomorphology 389, 107804. https://doi.org/10.1016/j.geomorph.2021.107804.
Joliffe, I., 2002. Principal Component Analysis. Springer. https://doi.org/10.1007/b98835.
Jordahl, K., Van den Bossche, J., Fleischmann, M., Wasserman, J., McBride, J., Gerard, J., Tratner, J., Perry, M., Garcia Badaracco, A., Farmer, C., Hjelle, G., Snow, A., Cochran, M., Gillies, S., Culbertson, L., Bartos, M., Eubank, N., Maxalbert, Bilogur, A., Rey, S., Ren, C., Arribas-Bel, D., Wasser, L., Wolf, L., Journois, M., Wilson, J., Greenhall, A., Holdgraf, C., Filipe, Leblanc, F., 2020. Geopandas/Geopandas: v0.8.1. https://doi.org/10.5281/zenodo.3946761.
Kassambara, A., Mundt, F., 2020. Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. URL: https://CRAN.R-project.org/package=factoextra.
Kavzoglu, T., Sahin, E.K., Colkesen, I., 2014. Landslide susceptibility mapping using GIS–based multi–criteria decision analysis, support vector machines, and logistic regression. Landslides 11, 425–439. https://doi.org/10.1007/s10346-013-0391-7.
Kingma, D., Ba, J., 2017. Adam: A method for stochastic optimization. https://doi.org/10.48550/arXiv.1412.6980.
Kirby, J., Grilli, S., Horrillo, J., Liu, P., Nicolsky, D., Abadie, S., Ataie-Ashtiani, B., Castro, M., Clous, L., Escalante, C., Fine, I., González-Vida, J.M., Løvholt, F., Lynett, P., Ma, G., Macías, J., Ortega, S., Shi, F., Yavari-Ramshe, S., Zhang, C., 2022. Validation and inter–comparison of models for landslide tsunami generation. Ocean Model 170, 101943. https://doi.org/10.1016/j.ocemod.2021.101943.
Kuhn, M., Silge, J., 2022. Tidy Modeling with R: A Framework for Modeling in the Tidyverse. Data Science, O'Reilly Media.
Kuhn, M., Wickham, H., 2020. Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles. URL: https://www.tidymodels.org.
Lee, S., 2019. Current and future status of GIS–based landslide susceptibility mapping: a literature review. Korean J. Remote Sens. 35, 179–193. https://doi.org/10.7780/kjrs.2019.35.1.12.
Lee, S., Evangelista, D.G., 2006. Earthquake–induced landslide–susceptibility mapping using an artificial neural network. Nat. Hazards Earth Syst. Sci. 6, 687–695. https://doi.org/10.5194/nhess-6-687-2006.
Lee, J.H., Sameen, M.I., Pradhan, B., Park, H.J., 2018. Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods. Geomorphology 303, 284–298. https://doi.org/10.1016/j.geomorph.2017.12.007.
Leoni, G., Barchiesi, F., Catallo, F., Dramis, F., Fubelli, G., Lucifora, S., Mattei, M., Pezzo, G., Puglisi, C., 2009. GIS methodology to assess landslide susceptibility: application to a river catchment of Central Italy. J. Maps 5, 87–93. https://doi.org/10.4113/jom.2009.1041.

Leung, W., Wong, J.C., Kwan, J.S., Petley, D.N., 2024. The use of digital technology for rock mass discontinuity mapping: review of benchmarking exercise. Bull. Eng. Geol. Environ. 83, 249. https://doi.org/10.1007/s10064-024-03730-w.
Liang, Z., Wang, C., Khan, K.U.J., 2021. Application and comparison of different ensemble learning machines combining with a novel sampling strategy for shallow landslide susceptibility mapping. Stoch. Env. Res. Risk A. 35, 1243–1256. https://doi.org/10.1007/s00477-020-01893-y.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22. URL: http://CRAN.R-project.org/doc/Rnews/.
Lima, P., Steger, S., Glade, T., Murillo-García, F.G., 2022. Literature review and bibliometric analysis on data-driven assessment of landslide susceptibility. J. Mt. Sci. 19, 1670–1698.
Lindgren, F., Rue, H., 2015. Bayesian spatial modelling with R-INLA. J. Stat. Softw. 63, 1–25. https://doi.org/10.18637/jss.v063.i19.
Liu, L.L., Zhang, J., Li, J.Z., Huang, F., Wang, L.C., 2022. A bibliometric analysis of the landslide susceptibility research (1999–2021). Geocarto Int. 37, 14309–14334. https://doi.org/10.1080/10106049.2022.2087753.
Liu, S., Wang, L., Zhang, W., He, Y., Pijush, S., 2023. A comprehensive review of machine learning–based methods in landslide susceptibility mapping. Geol. J. 58, 2283–2301. https://doi.org/10.1002/gj.4666.
Loche, M., Alvioli, M., Marchesini, I., Bakka, H., Lombardo, L., 2022. Landslide susceptibility maps of Italy: lesson learnt from dealing with multiple landslide types and the uneven spatial distribution of the national inventory. Earth Sci. Rev. 232, 104125 https://doi.org/10.1016/j.earscirev.2022.104125.
Loche, M., Alvioli, M., Marchesini, I., Lombardo, L., 2023. Landslide susceptibility within the binomial generalized additive model. In: EGU General Assembly 2023. https://doi.org/10.5194/egusphere-egu23-3566. Vienna, Austria, 24–28 April.
Lombardo, L., Mai, P., 2018. Presenting logistic regression–based landslide susceptibility results. Eng. Geol. 244, 14–24. https://doi.org/10.1016/j.enggeo.2018.07.019.
Lombardo, L., Tanyas, H., 2022. From scenario–based seismic hazard to scenario–based landslide hazard: fast–forwarding to the future via statistical simulations. Stoch. Env. Res. Risk A. 36, 2229–2242. https://doi.org/10.1007/s00477-021-02020-1.
Lucchese, L.V., de Oliveira, G.G., Pedrollo, O.C., 2021. Investigation of the influence of nonoccurrence sampling on landslide susceptibility assessment using artificial neural networks. Catena 198, 105067. https://doi.org/10.1016/j.catena.2020.105067.
Luckman, P.G., 1987. Slope Stability Assessment under Uncertainty: A First Order Stochastic Approach. University of California, Berkeley.
Luzi, L., Pergalani, F., Terlien, M., 2000. Slope vulnerability to earthquakes at subregional scale, using probabilistic techniques and geographic information systems. Eng. Geol. 58, 313–336. https://doi.org/10.1016/S0013-7952(00)00041-7.
Mărgărint, M., Grozavu, A., Patriche, C., 2013. Assessing the spatial variability of coefficients of landslide predictors in different regions of Romania using logistic regression. Nat. Hazards Earth Syst. Sci. 13, 3339–3355. https://doi.org/10.5194/nhess-13-3339-2013.
Marjanović, M., Kovačević, M., Bajat, B., Voženílek, V., 2011. Landslide susceptibility assessment using SVM machine learning algorithm. Eng. Geol. 123, 225–234. https://doi.org/10.1016/j.enggeo.2011.09.006.
McKinney, W., 2010. Data structures for statistical computing in Python. In: van der Walt, Stéfan, Millman, Jarrod (Eds.), Proceedings of the 9th Python in Science Conference, pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.
Mela, C., Kopalle, P., 2002. The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations. Appl. Econ. 34, 667–677. https://doi.org/10.1080/00036840110058482.
Merghadi, A., Yunus, A., Dou, J., Whiteley, J., ThaiPham, B., Bui, D., Avtar, R., Abderrahmane, B., 2020. Machine learning methods for landslide susceptibility studies: a comparative overview of algorithm performance. Earth Sci. Rev. 207, 103225 https://doi.org/10.1016/j.earscirev.2020.103225.
Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T., 2019. Importance of spatial predictor variable selection in machine learning applications - moving from data reproduction to spatial prediction. Ecol. Model. 411, 108815 https://doi.org/10.1016/j.ecolmodel.2019.108815.
Mirus, B., Woodard, J., 2023. Bayesian logistic regression and optimized XGBoost models for landslide susceptibility assessment. In: EGU General Assembly 2023. https://doi.org/10.5194/egusphere-egu23-11586. Vienna, Austria, 24–28 April.
Moreno, M., Steger, S., 2023. Slope unit size matters - why should the areal extent of slope units be considered in data–driven landslide susceptibility models?. In: EGU General Assembly 2023 https://doi.org/10.5194/egusphere-egu23-12943 Vienna, Austria, 24–28 April.
Moreno, M., Lombardo, L., Crespi, A., Zellner, P.J., Mair, V., Pittore, M., van Westen, C., Steger, S., 2024. Space–time data–driven modeling of precipitation–induced shallow landslides in South Tyrol, Italy. Sci. Total Environ. 912, 169166 https://doi.org/10.1016/j.scitotenv.2023.169166.
Openshaw, S., 1984. The modifiable areal unit problem. In: Concepts and Techniques in Modern Geography N. 38, Geo Books – Norwick.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, Édouard, 2011. Scikit–learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.
Petschko, H., Brenning, A., Bell, R., Goetz, J., Glade, T., 2014. Assessing the quality of landslide susceptibility maps–case study Lower Austria. Nat. Hazards Earth Syst. Sci. 14, 95–118. https://doi.org/10.5194/nhess-14-95-2014.
Pokharel, B., Alvioli, M., Lim, S., 2021. Assessment of earthquake–induced landslide inventories and susceptibility maps using slope unit–based logistic regression and geospatial statistics. Sci. Rep. 11, 21333 https://doi.org/10.1038/s41598-021-00780-y.
Prakash, N., Manconi, A., Loew, S., 2020. Mapping landslides on eo data: Performance of deep learning models vs. traditional machine learning models. Remote Sens. 12, 346. https://doi.org/10.3390/rs12030346.
R Core Team. R: A Language and Environment for Statistical Computing. URL: https://www.R-project.org.
Rabby, Y.W., Li, Y., Hilafu, H., 2023. An objective absence data sampling method for landslide susceptibility mapping. Sci. Rep. 13, 1740. https://doi.org/10.1038/s41598-023-28991-5.
Ranstam, J., Cook, J.A., 2018. Lasso regression. Br. J. Surg. 105, 1348. https://doi.org/10.1002/bjs.10895.
Regmi, N.R., Giardino, J.R., Vitek, J.D., 2010. Modeling susceptibility to landslides using the weight of evidence approach: Western Colorado, USA. Geomorphology 115, 172–187. https://doi.org/10.1016/j.geomorph.2009.10.002.
Reichenbach, P., Rossi, M., Malamud, B.D., 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.
Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., Müller, M., 2011. pROC: an open–source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12, 77. URL: http://expasy.org/tools/pROC/.
Rolain, S., Alvioli, M., Nguyen, Q., Nguyen, T., Jacobs, L., Kervyn, M., 2023. Influence of landslide inventory timespan and data selection on slope unit–based susceptibility models. Nat. Hazards 118, 2227–2244. https://doi.org/10.1007/s11069-023-06092-w.
Rossi, M., Reichenbach, P., 2016. LAND–SE: a software for statistically based landslide susceptibility zonation, version 1.0. Geosci. Model Dev. 9, 3533–3543. https://doi.org/10.5194/gmd-9-3533-2016.
Rossi, M., Bornaetxea, T., Reichenbach, P., 2022. LAND–SUITE V1.0: a suite of tools for statistically based landslide susceptibility zonation. Geosci. Model Dev. 15, 5651–5666. https://doi.org/10.5194/gmd-15-5651-2022.
Rumelhart, D., Hinton, G., Williams, R., 1986. Learning representations by back–propagating errors. Nature 323, 533–536. https://doi.org/10.1038/323533a0.
Sahin, E.K., 2020. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. SN Appl. Sci. 2, 1308. https://doi.org/10.1007/s42452-020-3060-1.
Samodra, G., Wahyudi, E.E., Susyanto, N., 2023. Cross validation technique preference for landslide susceptibility zoning based on slope unit and machine learning workflow. In: EGU General Assembly 2023. https://doi.org/10.5194/egusphere-egu23-11051. Vienna, Austria, 24–28 April.
Satyam, N., Abraham, M., Gupta, K., 2023. Resolution of data, type of inventory and data splitting in machine learning-based landslide susceptibility mapping. In: EGU General Assembly 2023. https://doi.org/10.5194/egusphere-egu23-4851. Vienna, Austria, 24–28 April.
Scaringi, G., Loche, M., 2023. Landslide susceptibility mapping via binomial generalized additive model. In: EGU General Assembly 2023. https://doi.org/10.5194/egusphere-egu23-6053. Vienna, Austria, 24–28 April.
Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., Toomet, O., Crowley, J., Hofmann, H., Wickham, H., 2022. R Package 'GGally': A Plotting System Based on the Grammar of Graphics. URL: https://github.com/ggobi/ggally.
Schlögel, R., Marchesini, I., Alvioli, M., Reichenbach, P., Rossi, M., Malet, J.P., 2018. Optimizing landslide susceptibility zonation: Effects of DEM spatial resolution and slope unit delineation on logistic regression models. Geomorphology 301, 10–20. https://doi.org/10.1016/j.geomorph.2017.10.018.
Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., Brenning, A., 2019. Hyperparameter tuning and performance assessment of statistical and machine–learning algorithms using spatial data. Ecol. Model. 406, 109–120. https://doi.org/10.1016/j.ecolmodel.2019.06.002.
Schreurs, G., Buiter, S., Boutelier, J., Burberry, C., Callot, J.P., Cavozzi, C., Cerca, M., Chen, J.H., Cristallini, E., Cruden, A., Cruz, L., Daniel, J.M., Da Poian, G., Garcia, V., Gomes, C., Grall, C., Guillot, Y., Guzmán, C., Hidayah, T., Hilley, G., Klinkmüller, M., Koyi, H., Lu, C.Y., Maillot, B., Meriaux, C., Nilfouroushan, F., Pan, C.C., Pillot, D., Portillo, R., Rosenau, M., Schellart, W., Schlische, R., Take, A., Vendeville, B., Vergnaud, M., Vettori, M., Wang, S.H., Withjack, M., Yagupsky, D., Yamada, Y., 2016. Benchmarking analogue models of brittle thrust wedges. J. Struct. Geol. 92, 116–139. https://doi.org/10.1016/j.jsg.2016.03.005.
Shano, L., Raghuvanshi, T., Meten, M., 2020. Landslide susceptibility evaluation and hazard zonation techniques – a review. Geoenviron. Disasters 7. https://doi.org/10.1186/s40677-020-00152-0.
Shcheglovitova, M., Anderson, R.P., 2013. Estimating optimal complexity for ecological niche models: a jackknife approach for species with small sample sizes. Ecol. Model. 269, 9–17. https://doi.org/10.1016/j.ecolmodel.2013.08.011.
Sin Yin, T., Othman, A., Chong Khoo, M., 2010. Dichotomous Logistic Regression with Leave–One–Out Validation. World Academy of Science, Engineering and Technology, pp. 538–547. https://doi.org/10.5281/zenodo.1057128.
Sinčić, M., Bernat Gazibara, S., Krkač, M., Lukačić, H., Mihalić Arbanas, S., 2023. A slope units based landslide susceptibility analyses using weight of evidence and random forest. In: EGU General Assembly 2023. https://doi.org/10.5194/egusphere-egu23-5755. Vienna, Austria, 24–28 April.
Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21, 3940–3941. https://doi.org/10.1093/bioinformatics/bti623.
Sirbu, F., 2023. Landslide susceptibility model based on random forest classification. In: EGU General Assembly 2023. https://doi.org/10.5194/egusphere-egu23-733. Vienna, Austria, 24–28 April.
Steger, S., Brenning, A., Bell, R., Petschko, H., Glade, T., 2016. Exploring discrepancies between quantitative validation results and the geomorphic plausibility of statistical

landslide susceptibility maps. Geomorphology 262, 8–23. https://doi.org/10.1016/j.geomorph.2016.03.015.
Steger, S., Schmaltz, E., Glade, T., 2020. The (f)utility to account for pre–failure topography in data–driven landslide susceptibility modelling. Geomorphology 354, 107041. https://doi.org/10.1016/j.geomorph.2020.107041.
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 https://doi.org/10.1016/j.scitotenv.2021.145935.
Sterlacchini, S., Ballabio, C., Blahut, J., Masetti, M., Sorichetta, A., 2011. Spatial agreement of predicted patterns in landslide susceptibility maps. Geomorphology 125, 51–61. https://doi.org/10.1016/j.geomorph.2010.09.004.
Süzen, M.L., Doyuran, V., 2004. Data driven bivariate landslide susceptibility assessment using geographical information systems: a method and application to Asarsuyu catchment, Turkey. Eng. Geol. 71, 303–321. https://doi.org/10.1016/S0013-7952(03)00143-1.
Thai Pham, B., Prakash, I., Dou, J., Singh, S., Phan Trong, T., Trung Tran, H., Minh Le, T., Van Phong, T., Dang Kim, K., Shirzadi, A., Tien Bui, D., 2020. A novel hybrid approach of landslide susceptibility modelling using rotation forest ensemble and different base classifiers. Geocarto Int. 35, 1267–1292. https://doi.org/10.1080/10106049.2018.1559885.
Tien Bui, D., Thai Pham, B., Quoc Nguyen, P., Hoang, N.D., 2016. Spatial prediction of rainfall–induced shallow landslides using hybrid integration approach of Least–Squares Support Vector Machines and differential evolution optimization: a case study in Central Vietnam. Int. J. Digit. Earth 9, 1077–1097. https://doi.org/10.1080/17538947.2016.1169561.
Torizin, J., Schüßler, N., 2023. Exploring the benchmark dataset for tasks related to landslide susceptibility assessment. In: EGU General Assembly 2023. https://doi.org/10.5194/egusphere-egu23-13362. Vienna, Austria, 24–28 April.
Trigila, A., Iadanza, C., Spizzichino, D., 2010. Quality assessment of the Italian landslide inventory using GIS processing. Landslides 7, 455–470. https://doi.org/10.1007/s10346-010-0213-0.
UCI, 2024. Machine learning repository. https://archive.ics.uci.edu/. Accessed: July 19, 2024.
Wang, N., Cheng, W., Marconcini, M., Bachofer, F., Liu, C., Xiong, J., Lombardo, L., 2022. Space–time susceptibility modeling of hydro-morphological processes at the chinese national scale. Eng. Geol. 301, 106586 https://doi.org/10.1016/j.enggeo.2022.106586.
Wang, T., Dahal, A., Fang, Z., van Westen, C., Yin, K., Lombardo, L., 2024. From spatio-temporal landslide susceptibility to landslide risk forecast. Geosci. Front. 15, 101765.
Wei, T., Simko, V., 2021. R Package 'Corrplot': Visualization of a Correlation Matrix. URL: https://github.com/taiyun/corrplot.
Wislocki, A., Bentley, S., 1991. An expert system for landslide hazard and risk assessment. Comput. Struct. 40, 169–172. https://doi.org/10.1016/0045-7949(91)90469-3.
Wood, S., 2017. Generalized Additive Models – An Introduction with R, 2nd edition. In: Mathematics & Statistics. Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279.
Wood, S.N., Augustin, N.H., 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecol. Model. 157, 157–177. https://doi.org/10.1016/S0304-3800(02)00193-X.
Yeon, Y.K., Han, J.G., Ryu, K.H., 2010. Landslide susceptibility mapping in Injae, Korea, using a decision tree. Eng. Geol. 116, 274–283. https://doi.org/10.1016/j.enggeo.2010.09.009.
Yesilnacar, E., Topal, T., 2005. Landslide susceptibility mapping: a comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). Eng. Geol. 79, 251–266. https://doi.org/10.1016/j.enggeo.2005.02.002.
Yong, C., Jinlong, D., Guo, F., Bin, T., Tao, Z., Hao, F., Li, W., Qinghua, Z., 2022. Review of landslide susceptibility assessment based on knowledge mapping. Stoch. Env. Res. Risk A. 36, 2399–2417. https://doi.org/10.1007/s00477-021-02165-z.
Zeng, T., Wu, L., Peduto, D., Glade, T., Hayakawa, Y.S., Yin, K., 2023. Ensemble learning framework for landslide susceptibility mapping: different basic classifier and ensemble strategy. Geosci. Front. 14, 101645 https://doi.org/10.1016/j.gsf.2023.101645.
Zuur, A.F., Ieno, E.N., Walker, N., Saveliev, A.A., Smith, G.M., 2009. Mixed Effects Models and Extensions in Ecology with R. Springer. https://doi.org/10.1007/978-0-387-87458-6.