



# A modular framework for FAIR shallow landslide susceptibility mapping based on machine learning

Ann-Kathrin Edrich<sup>1</sup> · Anil Yildiz<sup>1</sup> · Ribana Roscher<sup>2,3</sup> · Alexander Bast<sup>4,5</sup> · Frank Graf<sup>5,6</sup> · Julia Kowalski<sup>1</sup>

Received: 11 August 2023 / Accepted: 10 March 2024  
© The Author(s) 2024

## Abstract

Machine learning has grown in popularity in the past few years for susceptibility and hazard mapping tasks. Necessary steps for the generation of a susceptibility or hazard map are repeatedly implemented in new studies. We present a Random Forest classifier-based landslide susceptibility and hazard mapping framework to facilitate future mapping studies using machine learning. The framework, as a piece of software, follows the FAIR paradigm, and hence is set up as a transparent, reproducible and modularly extensible workflow. It contains pre-implemented steps from conceptualisation to map generation, such as the generation of input datasets. The framework can be applied to different areas of interest using different environmental features and is also flexible in terms of the desired scale and resolution of the final map. To demonstrate the functionality and validity of the framework, and to explore the challenges and limitations of Random Forest-based susceptibility and hazard mapping, we apply the framework to a test case. This test case conveys the influence of the training dataset on the generated susceptibility maps in terms of feature combination, influence of non-landslide instances and representativeness of the training data with respect to the area of interest. A comparison of the test case results with the literature shows that the framework works reliably. Furthermore, the results obtained in this study complement the findings of previous studies that demonstrate the sensitivity of the training process to the training data, particularly in terms of its representativeness.

**Keywords** Shallow landslides · Hazard mapping · Machine learning · Random Forest · FAIR data · Shallow landslide susceptibility

## 1 Introduction

Shallow landslides are destructive geohazards in mountainous areas that occur frequently and in large numbers after single triggering events. Therefore, they pose a threat to population and infrastructure. Natural trigger mechanisms that elicit enhanced shallow landslide activity are seismic events (Uchida et al. 2006) or strong subsurface water infiltration either through heavy precipitation or snow melt (Leonarduzzi et al. 2017). Shallowness refers to

failure planes located not more than 2–3 m below surface (Caine 1980; Rickli et al. 2019), with smaller mobilised masses compared to more deep-seated landslides.

The spatial distribution of susceptibility to shallow landslides can be effectively communicated through susceptibility maps, while their destructive potential is generally depicted in hazard maps. The planning of landslide hazard mitigation measures, such as soil nail walls (e.g. Maleki and Mir Mohammad Hosseini 2022) and soil bio- and eco-engineering measures (e.g. Bast et al. 2016; Graf et al. 2019), can be effectively guided using landslide susceptibility and hazard maps. Numerous studies have been conducted over the last few decades resulting in constantly evolving methods and approaches for landslide susceptibility and hazard mapping. Shano et al. (2020) provide an overview of approaches used in landslide susceptibility evaluation, including inventory-based prediction, expert evaluation, statistical, deterministic, probabilistic and distribution-free approaches. Selecting the most suitable approach, given a certain mapping task, depends on the landslide type to be mapped, the area of interest, map resolution and scale, data availability, and available resources including capability and skill set of the evaluator (Shano et al. 2020). Susceptibility and hazard maps differ in their definition as hazard includes information on the magnitude as well as a temporal dimension (Hervás and Bobrowsky 2009). Maps that are produced in this study are susceptibility maps indicating failure potential (1) or no failure potential (0). The assumptions, methods, advantages and limitations listed herein apply similarly to hazard maps.

Data-driven mapping based on machine learning (ML) algorithms can be applied whenever sufficient data on past events and environmental conditions are available. ML has taken an increasingly prominent role in shallow landslide hazard and susceptibility mapping in recent years (e.g. Shirzadi and Soliamani 2018; Pradhan and Kim 2020; Liu et al. 2021b). The resulting maps are promising and establish confidence in ML as a tool for hazard mitigation. A commonly used ML algorithm in landslide susceptibility and hazard mapping is the Random Forest (RF) (e.g. Stumpf and Kerle 2011; Zhang and Wu 2020; Liu et al. 2022; Dong et al. 2023). In comparison with other ML algorithms, it also yields satisfying results (e.g. Trigila et al. 2015; Chen et al. 2017; Sevgen et al. 2019; Karantanellis et al. 2021; Liu et al. 2021b; Youssef and Pourghasemi 2021; Feng and Guo 2023). The RF is a suitable algorithm for generating susceptibility maps due to its proven performance, its inherently increased explainability, and user friendliness.

The steps between susceptibility mapping-related research question and the generation of the final map are highly repetitive. Implementing tasks such as data preprocessing is time-consuming, and complex. While for other established mapping approaches, such as the ones based on physical models, pre-implemented frameworks, e.g. TRIGRS (Baum et al. 2002) exist, fewer literature is available for similar frameworks for data-based susceptibility and hazard mapping. Osona et al. (2014) introduced GeoFIS which allows the generation of maps based on expert opinion using the Mamdani fuzzy inference system. Sezer et al. (2017) presented a landslide susceptibility mapping module for the Netcad Architect Software. A Python-based add-on for the GIS software GRASS (Neteler et al. 2012) was presented by Bragagnolo et al. (2020). Sahin et al. (2020) developed a tool pack for landslide susceptibility mapping for R (R Core Team 2020) and ArcGis (Environmental Systems Research Institute, Inc. 2010a). Huang et al. (2022) introduced the SVM-LSM toolbox which is based on a support vector machine, allowing landslide susceptibility mapping which can be integrated into ArcGIS and ArcGIS Pro (Environmental Systems Research Institute, Inc. 2010a, b).

The available software packages differ in their approach, assumptions and capabilities. Most use secondary software such as GIS, which in some cases is even commercial.

Hence, in addition to this list, we introduce a generic Python-based (Python Software Foundation 2021) framework designed to facilitate future susceptibility and hazard mapping using RF that operates independently of secondary software. The framework aims to minimise the time and effort required by providing modular and flexible pre-implementations for repetitive preparatory steps. Furthermore, storing data pre-processing, input dataset generation, training and mapping parameters makes the results derived through the framework highly reproducible. It is also possible to return to previous mapping runs, remove or add e.g. features from or to the input datasets before rerunning the mapping step. This capability eliminates the need to run the entire framework from scratch. The framework enables the generation of binary event-based maps. These binary maps categorise susceptibility into two distinct classes: ‘susceptible’ and ‘not susceptible’. Such maps represent the fundamental form of a probabilistic susceptibility map.

The framework consists of three main components: (1) conceptualisation, (2) generation of the RF input datasets, including data pre-processing, and (3) map generation. Model validation is possible by checking the post-training dataset accuracy as well as scientific consistency by exploiting the RF’s intrinsic ability to provide feature importance information.

The challenge in constructing a generic framework lies in ensuring its flexibility and modularity to address diverse scientific questions. Key adjustable parameters encompass the scale of the area of interest, the data basis, and the resolution of the final map. It is crucial to emphasise that our aim is not to introduce a new susceptibility mapping approach, but rather to provide support for the streamlined application of the established RF approach.

When implementing the framework, great importance was attributed to:

- Reproducibility of the mapping result, extensibility in terms of included features and scalability of the area of interest. Prerequisites are the availability of suitable training data and sufficient computational power for the desired combination of scale and resolution of the final map.
- Explainability and transparency of the model and its results to strengthen the trust in the model and, as a consequence, also the reliability of the final map. This can be achieved through assessing scientific consistency, evaluating the validation dataset, as well as applying methods of Explainable Artificial Intelligence which have gained increasing importance in recent years. This includes for example the investigation of the feature importance of the trained model (see Sect. 3.7).
- FAIR-ness of the workflow and derived results. FAIR refers to findability, accessibility, interoperability and reusability primarily related to research data (Wilkinson and Dumontier 2016), but has been discussed as well for research software (Lamprecht and Garcia 2020).

Feasibility and applicability of the framework is demonstrated via a complementary test case focusing on shallow landslide susceptibility mapping in Switzerland. Flexibility and extensibility of the framework are exploited to investigate the influence of the training dataset’s composition on the mapping result. The obtained results are compared to findings of previous studies. This dual approach not only highlights the sensitivity of the ML approach to the provided data but, additionally underscores the validity and functionality of the framework.

Subsequently, this article has a twofold goal:

- Introducing a framework designed to facilitate future studies on landslide susceptibility and hazard mapping using RF. The framework offers pre-implemented solutions for repetitive tasks. At the same time, it showcases flexibility by allowing easy supplementation of modules tailored to individual research questions. This adaptability is possible due to the framework's modular structure and interoperability.
- Its application to a test case, allowing an investigation of the impact of the training dataset on both model and mapping accuracy. These results serve to validate the framework's output by comparison to literature and to highlight the sensitivity of the underlying approach to its input data.

## 2 Framework for shallow landslide susceptibility and hazard mapping

### 2.1 Machine learning in landslide susceptibility mapping

ML has taken a more prominent role in susceptibility and hazard mapping over the last years (Shirzadi and Soliamani 2018; Liu et al. 2023) due to increased computational capability and the increased availability of datasets—especially publicly-available open-access data generated in research projects or published by governmental bodies. The core principle of ML-based susceptibility and hazard mapping is assuming that future landslides will occur under similar conditions as historic landslides (Tien Bui et al. 2016). ML is applied to identify patterns among conditions, parameters, and statistics of past landslides. Subsequently, the locations within the area of interest are examined for the occurrence of these patterns. This process allows conclusions about their potential susceptibility to failure. The information provided to the model is a collection of prevailing environmental conditions at sites of historic landslides (presence data), and at sites where no landslides were documented in the past (absence data). It is thereby inherently assumed that those conditions are static, meaning the conditions depicted in the datasets are the same as at the date of occurrence (Reichenbach et al. 2018). The simplest form of a probabilistic map is a binary map classifying susceptibility in the form of 'susceptible' or 'not susceptible'. This type of map is commonly extended by defining various probability classes, represented by a colour-code indicating the susceptibility level in different zones (Kavzoglu et al. 2019; Du et al. 2020; Zhang and Wu 2020). The maps maintain their validity and reliability only as long as the incorporated information in the form of underlying features is accurate.

### 2.2 Data collection and pre-processing

Data involved in landslide susceptibility and hazard mapping typically is inventory data documenting past landslides, environmental data, representing the state and characteristics of the slopes, and data on triggering factors (Shano et al. 2020). The quality of any output from data-driven susceptibility mapping approaches relies heavily on the quality and properties of the integrated datasets (Lee et al. 2004; Gaidzik and Ramírez-Herrera 2021), which are often collected from a variety of third-party sources with varying file formats, coordinate reference systems, spatial extents and resolutions. A homogenisation effort is needed prior to any mapping in order to ease the generation of the input datasets and increase the reproducibility. Examples of such efforts are the application of interpolation algorithms to achieve a uniform resolution, coordinate transformation to unify the coordinate reference system, and cropping to a matching spatial extent.

For the test case, the datasets in Sect. 2.2.2 were collected, cropped and interpolated to a 25 m resolution using the preprocessing implemented in the framework. Necessary transformations to the same coordinate reference system were conducted in the Geographic Information System QGIS (QGIS Development Team 2020). We chose only publicly available datasets because the ability to reproduce and replicate a susceptibility or hazard assessment is crucial for its reliability. An overview of all collected and included datasets can be found in Table 1.

### 2.2.1 Landslide inventory

For the test case, we utilise the *Hangmuren* database published by the Swiss Federal Institute for Forest, Snow and Landscape Research WSL (Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Research Unit Mountain Hydrology and Mass Movements 2023a). The database comprises 759 entries recorded on 13 different dates between 1997 and 2014 (status March 2021) through field campaigns following heavy rainfall events in Switzerland (Rickli et al. 2019; Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Research Unit Mountain Hydrology and Mass Movements 2023a). Entries lacking either location information or a timestamp were excluded. Additionally, data before 2000 was omitted since precipitation data is available since June 2000.

The remaining 476 landslides exhibit a clustered rather than even distribution (see Fig. 1a), possibly due to the nature of the database compilation. The information extracted from the *Hangmuren* database is limited to the location and date of landslide occurrence. Further information on the subset of the *Hangmuren* database used in this study is provided in Table 2. Henceforth, the entries of the *Hangmuren* database will be referred to as presence locations, in contrast to the absence locations introduced in Sect. 2.4.

### 2.2.2 Geospatial datasets

Most studies on landslide susceptibility and hazard mapping, independent of their choice of methodology, use similar predictor variables describing the prevailing environmental conditions from the domains of topography, land cover, slope hydrology, soil and geological properties, sometimes also anthropology and meteorology (Bui et al. 2012; Conforti et al. 2014; Kumar et al. 2017; Dang et al. 2019; Dou and Yunus 2019; Liu et al. 2021b; Stanley et al. 2021). The selection of input features depends on factors such as the area of interest, the availability of respective datasets, and their overall quality. Integrating environmental features into the mapping process inherently assumes that they accurately describe the environmental conditions of the past as well (Reichenbach et al. 2018).

For the test case scenario presented in Sect. 3, we selected features that physically influence the stability of the ground.

Topographic features are among the most essential features for landslide susceptibility and hazard mapping and are therefore used in almost every study (e.g. Conforti et al. 2014; Pandey and Sharma 2017; Kuradusenge et al. 2020; Liu et al. 2021b). Typically, elevation and geomorphological parameters such as slope angle and aspect are extracted or derived from digital elevation models. Elevation is associated with the changing composition and nature of the subsurface, and slope parameters reasonably affect the likelihood of slope failures.

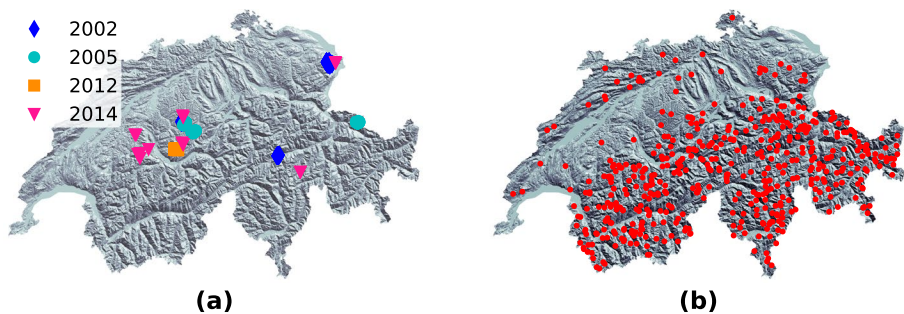
Strong infiltration of water into the ground can increase the probability of the occurrence of shallow landslides (Caine 1980; Wang et al. 2020). This infiltration is most

**Table 1** Overview of the included datasets in the test case scenario either as features in the RF classifier input datasets or in the generation of the absence location sampling masks

Parameter	Unit	Dataset	Source	Resolution (m)	Extent
Elevation	m	DHM25	swisstopo	25	Switzerland
Slope	degrees				
Aspect	-				
Sand	%	Topsoil Physical Properties for Europe	ESDAC	500	Europe
Silt					
Clay					
Bulk density	g cm <sup>-3</sup>				
Coarse fragments	%				
USDA classes	Categorical				
Avail. water capacity	-				
Saturated water content	cm <sup>3</sup> cm <sup>-3</sup>	3D Soil Hydraulic Database of Europe	ESDAC	250	Europe
$\alpha$ (SWRC)	cm <sup>-1</sup>				
$n$ (SWRC)	-				
Land cover	Categorical	Corine Land Cover 2018, Version 2020_20u1	Copernicus Land Monitoring Service	100	Europe
Tree cover density	%	High Resolution Layer: Tree Cover Density 2018		10	Europe
Imperviousness	%	High Resolution Layer: Imperviousness Density 2018		10	Europe
Water and wetness	%	High Resolution Layer: Wetness 2018		10	Europe
Precipitation	mm	IMERG Final run Global Precipitation measurement mission	NASA	10,000	Global

**Table 2** Summary of the subset of the *Hangmuren* database used for the study (as of March 2021)

Date (yyyy-mm-dd)	Number of records
2002-09-01	106
2002-07-16	64
2002-11-14	1
2002-11-15	34
2005-08-22/23	217
2012-07-04	38
2014-07-12	2
2014-07-24	4
2014-07-28	5
2014-08-11	3
2014-11-05	2



**Fig. 1** **a** Distribution of the presence data given in the *Hangmuren* database by the Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Research Unit Mountain Hydrology and Mass Movements (2023a), colour-coded by their year of occurrence (status March 2021), **b** distribution of the sampled absence locations (red dots). Basemaps derived from the DHM25 (Federal Office of Topography swisstopo 2005)

commonly caused by heavy rain events but can also be caused by, e.g. snow melt. Previous studies have also shown that not necessarily the precipitation on the event date itself but rather the accumulation of precipitation over the antecedent days increases the likelihood of an event significantly (Kuradusenge et al. 2020). Conditions such as soil type or bulk density play a crucial role in controlling the ability of water to infiltrate into the ground. Consequently, soil-related information is often incorporated into the mapping process.

Land cover information supports the RF decision by providing information on which types of land cover are prone to the occurrence of landslides.

The digital elevation model utilised for the test case is the DHM25, with a resolution of 25 m, provided by the Federal Office of Topography swisstopo (2005). Slope angle and aspect maps were derived using QGIS.

Soil-related features were extracted from the Topsoil Physical Properties for Europe dataset which is derived from the LUCAS topsoil data (Ballabio et al. 2016). The downloaded raster files with a resolution of 500 m are based on the LUCAS point-data. Information contained are percentage of coarse fragments, bulk density, available water

capacity, the soil type classification according to the United States Department of Agriculture (USDA) and sand, silt and clay content.

Parameters of the soil water retention curve (SWRC) were extracted from the 3D Soil Hydraulic Database of Europe, which is available with a resolution of 250 m (Tóth et al. 2017).

Precipitation data covering the period from June 2000 to December 2020 was obtained from NASA's Global Precipitation Measurement mission (Huffman et al. 2019). Precipitation was incorporated in the form of the maximum precipitation that was observed within this 20 year period at each considered location.

Land cover information has been derived from the Corine Land cover dataset, which has a resolution of 100 m and distinguishes 44 land cover classes (Copernicus Land Monitoring Service 2018a). Vegetation-related information is integrated through the tree cover density dataset with a resolution of 10 m (Copernicus Land Monitoring Service 2018b).

### 2.3 Input datasets for the Random Forest

The input datasets for the RF consist of the training, validation and prediction dataset, all having a tabular structure. Each row represents one presence or absence location and each column one feature, i.e. environmental parameter. The validation dataset is split from the training dataset before it is used for training with a ratio of 75:25 since there is no universal agreement on the ratio of training to validation data (Nurwati et al. 2022; Saha et al. 2021). The validation dataset is independent and unknown to the RF and can therefore be used to evaluate the accuracy of the trained model (see Sect. 3.6). The prediction dataset contains the same set of features as the training dataset for all locations within the area of interest.

### 2.4 Absence locations

Presence data offer insights into the conditions at locations of past landslides, while absence data provide information on locations where it can be assumed that no landslide took place in the past. The RF classifier requires both types of data for prediction (Belgiu and Drăguț 2016).

There is an abundance of possible locations to sample absence locations from as the occurrence of landslides is still comparably rare and spatially limited. However, landslide inventories in general but especially also event-based inventories as the one used in this study have to be assumed to be incomplete, meaning that a lot of landslides especially small, older or remote ones are not captured. This introduces uncertainty into the training dataset due to the possibility of sampling absence locations at locations that should be classified as presence locations. The choice of absence locations should also be meaningful with regard to the area of interest and the presence locations.

Random absence locations sampling, sometimes using a buffer zone around the recorded historic landslides, is one of the most common approaches (e.g. Taalab et al. 2018; Hong et al. 2019). More systematic approaches to tackle this task by choosing the absence data in the feature space have been proposed by Xiao et al. (2010) and Zhu et al. (2019).

Previous studies chose a wide range of ratios between the number of presence and absence locations. Stanley et al. (2021) for example use 9700 presence locations and over 1 million absence locations for a global nowcasting model while the most common approach (e.g. Regmi et al. 2014; Taalab et al. 2018; Dang et al. 2019), chooses an equal



number of presence and absence locations. The choice of ratio is individual to the conducted study though the ratio of 1:1 has been shown to be well suited in previous studies and presented by Hong et al. (2019) as the preferred ratio to prevent underestimation of landslide susceptibility.

Therefore, for the test case, a ratio of 1:1 was chosen and the influence of the ratio was investigated in Sect. 3.3. To ensure meaningful sampling, absence locations are randomly selected based on specific criteria: they must not be within 50 m of a landslide location, should be away from water bodies and sealed locations, and must have a slope angle between 20° and 50°. These restrictions were made based on domain knowledge. Figure 2a shows the areas among which the absence locations were sampled and Fig. 1b their distribution.

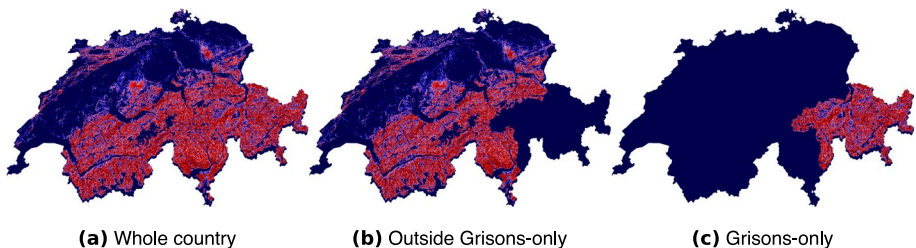
## 2.5 Technical implementation

Susceptibility and hazard maps are derived through three main steps: (1) conceptualisation, (2) generation of the RF input datasets, including data pre-processing, and (3) map generation. Validation of the inputs and results at each step is essential to ensure reliability and trust.

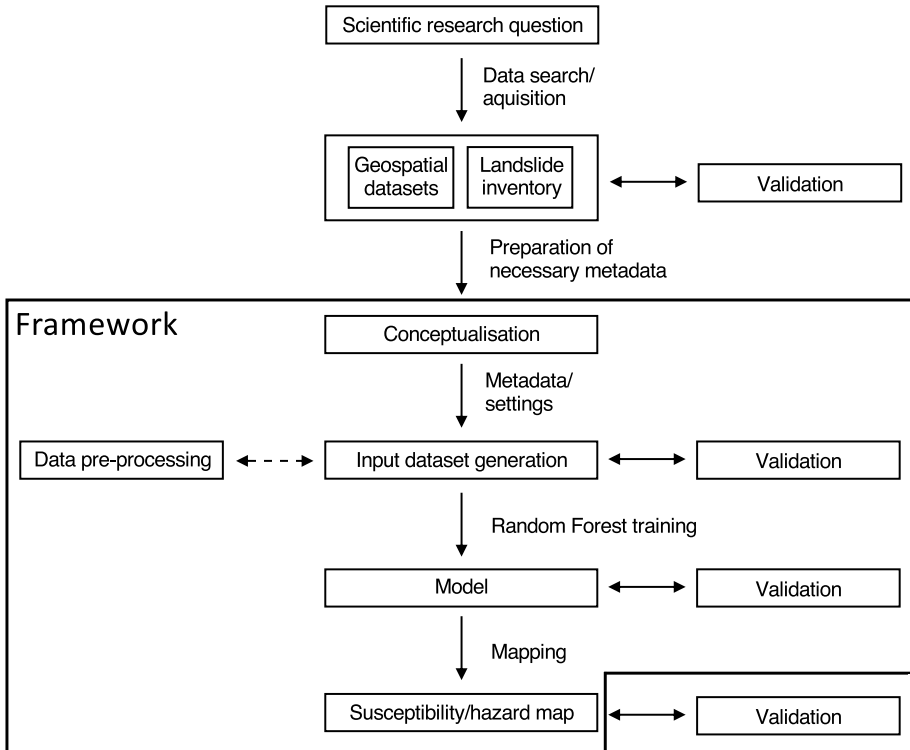
We established a Python-based framework that offers a generic yet flexible pre-implementation of these steps (Fig. 3). During the conceptualisation, the properties of the desired map are defined, such as resolution and area of interest. Geodata is pre-processed and prepared as input datasets. Pre-processing includes especially cropping to a uniform extent and interpolation to a uniform resolution. It has been ensured to account also for different kinds of datasets with their individual heterogeneous properties. Through parallelisation and size- and resolution-sensitive pre-processing, the framework is configured to operate on local machines. The map generation is based on a RF classifier (see Sect. 2.6).

Extensibility of the framework is ensured through easy generation of the input datasets including addition and removal of features and presence/absence instances. Training and prediction properties are stored and can be accessed also at a later point to support reproducibility of results.

The RF was set up using the python package sklearn (Pedregosa and Varoquaux 2011; Hao and Ho 2019). The output is a binary probabilistic map (see Sect. 2.1).



**Fig. 2** Illustration of the three sampling areas that were defined to investigate the influence of the sampling strategy on the mapping result in experiment 2. Red areas indicate possible sampling locations, in blue areas no absence data is sampled, **a** Sampling all over Switzerland, **b** sampling only outside Grisons, **c** sampling only within Grisons



**Fig. 3** Graphical illustration of the steps included in the presented landslide susceptibility and hazard mapping framework

## 2.6 Random Forest classifier

The RF is an ensemble tree algorithm consisting of an ensemble of individual decision trees. It can be configured either as a classifier, where it assigns a class to a location (e.g. ‘susceptible’ or ‘not susceptible’ in the context of susceptibility and hazard mapping), or as a regression model, where it assigns a numeric value, such as a factor of safety (Cutler et al. 2012). In the case of a classifier, the ultimate decision of the RF is determined by the majority vote of individual decision trees. In the regression case, the outcome of all trees is averaged (Breiman 2001).

The RF is computationally less expensive compared to other ensemble tree methods, making it suitable for high-dimensional or large-scale problems. The RF is a non-linear, non-parametric algorithm, allowing to account for complex interactions and non-linearities among the input variables. It can deal with large datasets of both continuous and categorical input features, and its implementation is straightforward as it does not require an extensive hyperparameter tuning. (Hastie et al. 2009; Cutler et al. 2012; Taalab et al. 2018). A RF classifier of 100 trees and depth of 20 nodes is used in this study.

### 3 Test case: Grisons

#### 3.1 Scenario

A test case scenario was defined to demonstrate the application of the framework and to conduct computational experiments to investigate the influence of the training dataset onto the RF-based susceptibility mapping result. Rainfall-triggered shallow landslide susceptibility in canton Grisons, Switzerland, was chosen as test case scenario due to (1) the mountainous terrain, which increases the likelihood of the occurrence of shallow landslides, and (2) the availability of high quality and high-resolution open-access geodata for Switzerland. The output of the workflow for this test case is a binary susceptibility map with a resolution of 25 m. Features included in the mapping process are elevation, slope angle, aspect, sand, silt and clay content, bulk density, percentage of coarse fragments, USDA classes, available water capacity, parameters saturated water content,  $\alpha$  and  $n$  of the SWRC, land cover, tree cover density and maximum observed precipitation (see Table 1). Pre-processing of the geospatial datasets focuses on cropping to the same extent and interpolating to the same resolution equal to the desired resolution of the final susceptibility map.

#### 3.2 Experiment 1: Feature selection

Although studies typically integrate similar features, the influence of the combination of features in the training dataset is rarely considered.

Eight feature subsets were generated from the full set of available features to investigate the influence of their combination (Table 3). Subsets 1 and 2 contain randomly selected features among all environmental domains, subsets 3 – 6 contain only features of the same kind of environmental factors. The features in subsets 7 and 8 are chosen according to their importance for the reference RF model (see Sect. 3.7). A new RF model was trained with each subset and consequently used for susceptibility mapping.

#### 3.3 Experiment 2: Absence locations sampling strategies

The influence of the sampling strategy for absence locations has rarely been discussed in literature (Hong et al. 2019; Lima et al. 2022). To investigate the influence of the choice of sampling area, and the importance of the ratio of number of presence locations to absence locations, different sampling strategies were defined. We defined three sampling areas: the whole of Switzerland, only within the area of interest, and the whole of Switzerland excluding the area of interest (Fig. 2). In these areas, we sampled absence locations based on the criteria presented in Sect. 2.4. Subsequently, these absence locations were utilised to generate training datasets, which were then combined with the 476 presence locations in the following ratios: 476:100, 476:238, 476:476, 476:700, 476:945, 476:1500 (Table 4). These training datasets were then used to create, compare, and evaluate susceptibility maps.

**Table 3** Overview of the feature subsets in experiment 1 as well as the validation results of the trained model and generated susceptibility maps using the SSLI and validation dataset. The rating system using the SSLI is described in Sect. 3.6. For the validation dataset the number of incorrect predictions out of the total number of instances is provided

Subset	Content	Features	Rating SSLI, validation dataset	Size of suscep. area (km <sup>2</sup> )
1	Randomly selected	Maximum precipitation Slope USDA classes Bulk density Tree cover density	202/25/82/485, 13/238, 5%	781
2	Randomly selected	Clay $\alpha$ (SWRC) Elevation Coarse fragments Land cover	277/47/57/413, 17/238, 7%	594
3	Soil-only	USDA classes Sand Silt Clay Bulk density Coarse fragments	372/27/37/358, 28/238, 12%	477
4	Land cover-only	Land cover Tree cover density	77/26/117/574, 52/238, 22%	1570
5	Hydrology-only	Maximum precipitation Avail. water capacity, $n$ (SWRC), $\alpha$ (SWRC), Saturated water content	437/21/33/303, 14/238, 6%	599
6	Topography-only	Slope Elevation Aspect	57/29/138/570, 49/238, 21%	1126
7	Most important features	Maximum precipitation Elevation Bulk density Land cover Coarse fragments $\alpha$ (SWRC) Saturated water content Slope	283/34/42/435, 9/238, 4%	526

**Table 3** (continued)

Subset	Content	Features	Rating SSLI, validation dataset	Size of suscep. area (km <sup>2</sup> )
8	Least important features	Aspect <i>n</i> (SWRC) USDA classes Sand Avail. water capacity Clay Tree cover density Silt	447/38/53/256, 16/238, 7%	379

### 3.4 Experiment 3: Representativeness of the training dataset

As has been described in Sect. 2.2.1 and can be seen in Fig. 1a, the presence locations are highly clustered due to the event-based characteristic of the inventory. In susceptibility and hazard mapping, some studies use only landslides of a single triggering event within the scope of a small area as training data (e.g. Yeon et al. 2010; Trigila et al. 2015; Dou and Yunus 2019; Liu et al. 2021b). For larger areas of interest, it has to be assumed that in order to ensure that training data is representative, a wider range of information is necessary to capture the possible variability of the environmental conditions. To illustrate the influence of the availability of limited information on the mapping result, scenarios were investigated where only landslides of a certain triggering event or limited time frame were included in the training and validation dataset. We ensured that the ratio of presence to absence locations in the training dataset is still 1:1 to be comparable also to the full training dataset reference map (Fig. 4). Table 5 sums up the generated presence data subsets.

### 3.5 Results

A susceptibility map was produced for the Swiss canton Grisons (Fig. 4) using the framework with the implementation and data described in Sect. 2. Particularly, the north of the canton is mapped as susceptible to shallow landslide occurrence. This map serves as a reference for the evaluation of all computational experiments described above.

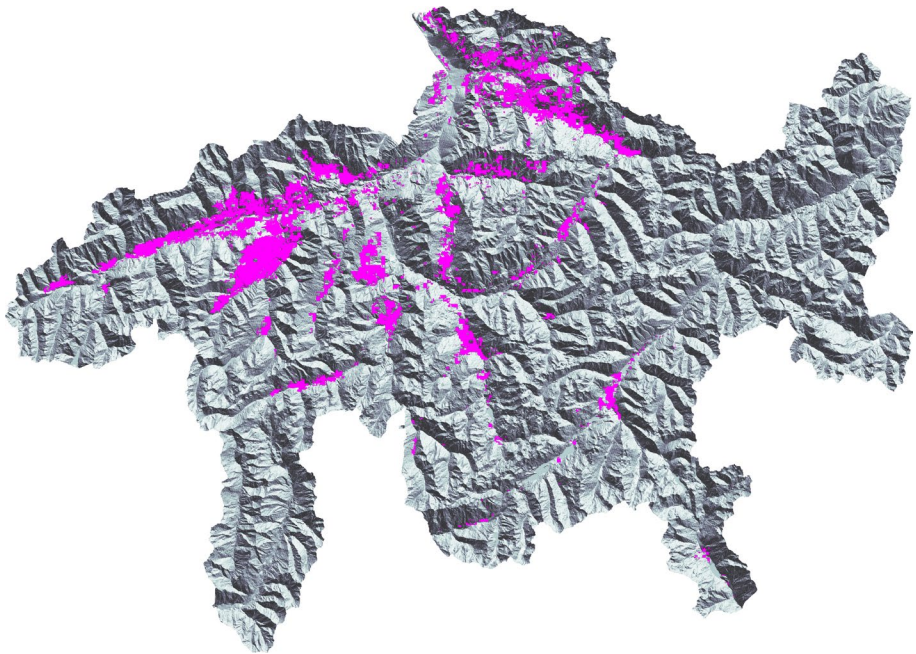
Figure 5 shows the susceptibility maps that were generated in the scope of experiment 1. Table 3 provides an overview of all variations of the training dataset, the size of the susceptible area for the individual maps as well as the results of the applied validation strategies (Sect. 3.6). The resulting maps differ visually significantly from one another and also in the total area mapped as susceptible to landslide occurrence. Especially, the extreme cases in terms of feature selection (subsets 3–6, see Fig. 5a–d) show a strong discrepancy to the reference map in Fig. 4. For subsets 3 and 5 in Fig. 5b, a, the size of the area mapped as susceptible is comparable to the reference map. In contrast, using only topography and

**Table 4** Overview of the sampling strategies used to investigate the influence of the sampling area and the ratio of the number of presence to absence locations on the prediction result in experiment 2, as well as the validation results using the SSLI and the validation dataset. The rating system using the SSLI is described in Sect. 3.6. For the validation dataset the number of incorrect predictions out of the total number of instances is provided

Strategy	Sampling area	Presence:absence locations	Size of suscep. area (km <sup>2</sup> )	Rating SSLI [0/1/2/3]	Validation dataset
1	Whole country with restrictions	476:100	1197	126/932/627	7/144, 5%
		476:238	707	220/15/40/519	7/179, 4%
		476:476	454	273/33/56/432	9/239, 4%
		476:700	285	348/44/53/349	13/294, 4%
		476:945	193	416/40/66/272	15/356, 4%
2	Only outside Grisons	476:1500	123	504/39/55/196	8/494, 2%
		476:100	1736	86/14/31/663	8/144, 6%
		476:238	943	204/21/47/522	15/179, 8%
		476:476	585	261/38/58/437	19/239, 8%
		476:700	403	340/29/51/374	16/294, 5%
3	Only within Grisons	476:945	339	352/45/82/315	19/356, 5%
		476:1500	234	417/65/93/219	19/494, 4%
		476:100	936	152/27/29/586	4/140, 3%
		476:238	723	210/15/41/528	6/179, 3%
		476:476	503	288/24/40/442	9/238, 4%
		476:700	352	336/40/50/368	9/294, 3%
		476:945	236	433/39/44/278	11/356, 3%
		476:1500	176	517/26/38/213	9/494, 2%

**Table 5** Overview of the test cases performed to investigate the influence of the representativeness of the training dataset on the prediction result in experiment 3, as well as the validation results using the SSLI and the validation dataset. The rating system using the SSLI is described in Sect. 3.6 and for the validation dataset the number of incorrect predictions out of the total number of instances is provided

Subset	Number of events	Date [yyyy(-mm-dd)]	Size of suscep. area (km <sup>2</sup> )	Rating SSLI [0/1/2/3]	Validation dataset
1	205	2002	188	574/18/20/182	9/103, 9%
2	106	2002-09-01	138	701/14/16/63	2/53, 4%
3	64	2002-07-16	40	710/21/17/46	1/32, 3%
4	217	2005-08-22/23	261	573/7/21/193	1/109, 1%

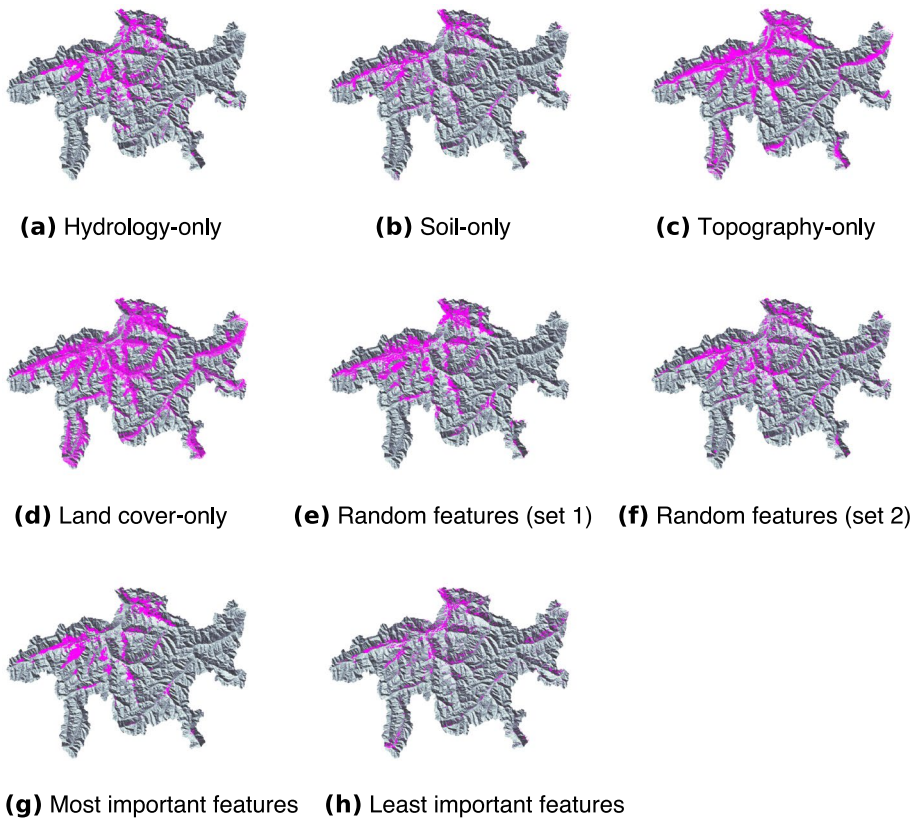


**Fig. 4** Shallow landslide susceptibility map for the canton Grisons, Switzerland. Purple areas indicate susceptibility to failure. The map serves as reference for assessing the maps generated in the context of the test case. Basemap derived from the DHM25 (Federal Office of Topography swisstopo 2005)

land cover-related features (subsets 6 and 4, see Fig. 5c, d), compared to the reference, leads to a strong increase of the area mapped as susceptible.

The comparison between subsets 1 and 2 (Fig. 5e, f) reveals interesting insights as even though both sets contain parameters from different environmental domains, the mapping results vary strongly. The susceptible area for subset 1 is notably larger than that of subset 2. At the same time, subset 2 maps areas as susceptible that subset 1 does not capture.

Figure 5g, which is derived from subset 7, only shows small discrepancies to the reference map (Fig. 4). Using only less important features (Fig. 5h) also results in areas mapped as susceptible that are not mapped in Fig. 5g, respectively the reference map.



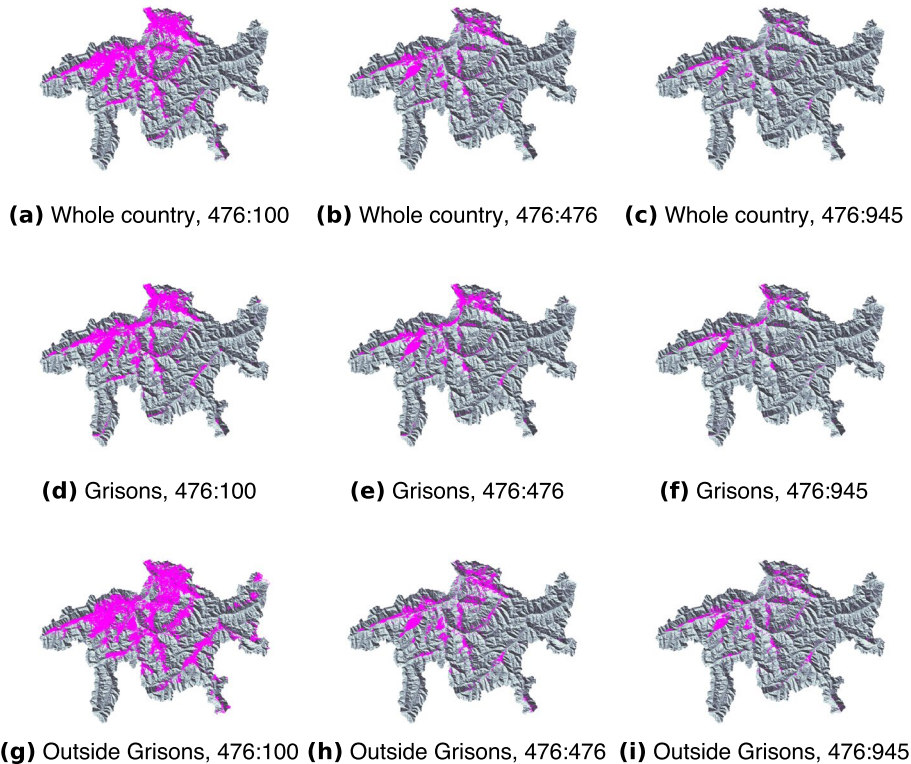
**Fig. 5** Susceptibility maps generated from the training data subsets given in Table 3 for experiment 1. Purple areas indicate susceptibility to shallow landslide occurrence, **a** Hydrology-only, **b** Soil-only, **c** Topography-only, **d** Land cover-only, **e** Random features (set 1), **f** Random features (set 2), **g** Most important features, **h** Least important features. Basemaps derived from the DHM25 (Federal Office of Topography swisstopo 2005)

Figure 6 shows exemplary susceptibility maps for the canton Grisons generated in the scope of experiment 2. Table 4 provides an overview on the tested variations of the training dataset, the size of the area classified as susceptible as well as the results of the applied validation strategies. The ratio of the number of presence to absence locations directly impacts the total area mapped as susceptible. The fewer absence locations are included in the mapping, the larger the susceptible area.

Changing the sampling strategy does not change the general area in which susceptibility is mapped. The effect of the ratio on the map is significantly larger than the choice of sampling area.

In the scope of experiment 3 susceptibility maps were created based on a subset of the presence and absence locations of the training dataset (see Fig. 7) using the methodology described in Sect. 3.4. The mapping results vary notably among the different subsets and also differ from the reference map. Table 5 offers an overview of the experiment, the results, and the outcome of the validation.





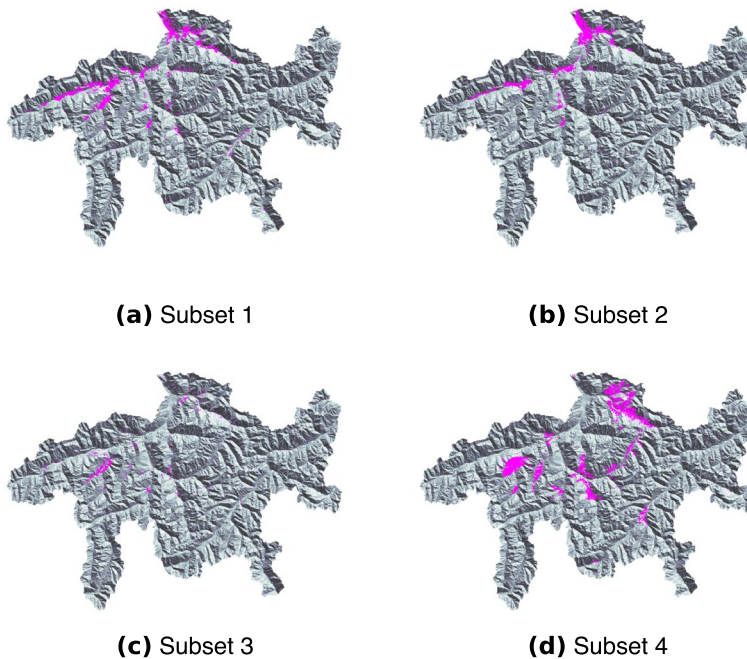
**Fig. 6** Susceptibility maps generated using different absence location sampling strategies as described in Table 4 in the context of experiment 2. Purple areas indicate susceptible areas. Shown are maps varying in sampling area and number of presence locations compared to number of absence locations, **a** Whole country, 476:100, **b** Whole country, 476:476, **c** Whole country, 476:945, **d** Grisons, 476:100, **e** Grisons, 476:476, **f** Grisons, 476:945, **g** Outside Grisons, 476:100, **h** Outside Grisons, 476:476, **i** Outside Grisons, 476:945. Basemaps derived from the DHM25 (Federal Office of Topography swisstopo 2005)

### 3.6 Validation

The quality of the maps produced in the test case is assessed in three different ways: (1) the uncertainties in the landslide inventory are discussed, (2) the accuracy of the trained RF models and the resulting susceptibility maps is investigated using an independent landslide inventory as well as the validation dataset, and (3) the feature importance is assessed. Together, these three approaches provide a sufficient overview of the quality of the models and the resulting maps.

#### 3.6.1 Uncertainties in the datasets

The quality of the input datasets significantly influences the quality of the mapping result (Thiery et al. 2020). Uncertainties in geospatial input data occur due to inaccuracy, imprecision, ambiguity, vagueness, subjectivity, or other unknown and non-quantified errors (Kinkeldey et al. 2017). They can be introduced at any point along the pipeline, from



**Fig. 7** Susceptibility maps based on the subsets of the training dataset presented in Table 5 for experiment 3. Purple areas indicate susceptibility to shallow landslide occurrence. Maps were generated using the following subsets, **a** Subset 1, **b** Subset 2, **c** Subset 3, **d** Subset 4. Basemaps derived from the DHM25 (Federal Office of Topography swisstopo 2005)

raw data acquisition to the incorporation as features in the ML input dataset. Information on uncertainties is typically not provided along with the dataset, and therefore hard to estimate.

Due to its vital importance for the present study, a brief assessment of uncertainties in the *Hangmuren* database is presented herein. The database was assembled through field work (Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Research Unit Mountain Hydrology and Mass Movements 2023a), and therefore inaccuracies related to position determination apply (Abraham et al. 2021) which is especially important with respect to the aspired high resolution of the susceptibility map. The distribution of the recorded landslides is clustered (see Fig. 1a) and therefore reduced representativeness has to be considered when interpreting. As this study aims at presenting a mapping framework including its application to a test case scenario and therefore is of theoretical nature, this is of reduced importance.

### 3.6.2 Model and susceptibility map validation

A typical means of assessing the quality of a ML model is the evaluation of the validation (also called test) dataset (Arora et al. 2004; Van Den Eeckhaut et al. 2009; Dou et al. 2015). The output of the RF at the locations of the presence and absence locations in the validation dataset are compared with the known correct classifications. From the number of correctly predicted instances, the accuracy is calculated. A small or not representative

validation dataset reduces the meaningfulness of the evaluation. The validation dataset of the present study has a size of 238 entries (25% of all absence and presence data). For the reference model, in total nine entries of the validation dataset were predicted incorrectly (ca. 4%).

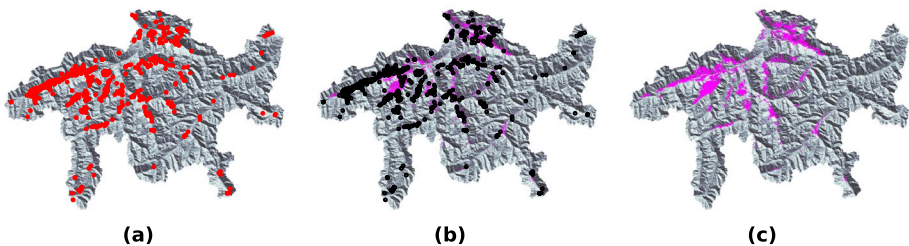
The entries in the *Hangmuren* database are clustered, resulting in similarities between the entries of the validation dataset and the training dataset. Therefore, to complement the validation dataset, the same approach was applied to the modified *Supplemented Swiss Landslide Inventory* (SSLI) (Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Research Unit Mountain Hydrology and Mass Movements 2023b; Bebi et al. 2019). The SSLI contains events that are less clustered and cover also regions where no landslides were recorded in the *Hangmuren* database. It was not used for training as it is not publicly-available, but its availability allows further assessment of the RF output.

Figure 8 shows the distribution of the landslides recorded in the SSLI database as well as their location on top of the areas classified as susceptible in the reference map. A visual comparison between the location of the SSLI entries and the susceptible areas reveals a good match. Apart from a few examples, e.g. in the south-west and north-east of the canton, most SSLI entries are located in good accordance with susceptible areas.

To be able to quantify the accuracy of the susceptibility analysis with respect to the SSLI, a scoring system is used. If the location of an entry in the SSLI matches a location classified as susceptible, the score 3 is assigned. Respectively, a score 2 or 1 are assigned if the location of the SSLI entry is one or two pixels away from a pixel marked as susceptible. Otherwise, a score 0 is assigned. This pattern can be applied to all entries in the SSLI within the area of interest of this study to get an overview of the mapping accuracy. In total, the SSLI contains 794 landslides in Grisons. Out of these, 432 get a score of 3, 56 a score of 2, 33 a score of 1 and 273 a score of 0 for the reference prediction which equals an accuracy of 66% taking the scores  $\neq 0$  into account.

Table 3 summarizes the results of the comparison of the susceptibility maps generated for experiment 1 with the SSLI. Judging from the numbers, the only-land cover data map is most accurate but Fig. 5 and the total susceptible area need to be taken into account as well which shows the likely massive overestimation of susceptible areas.

Table 4 shows the equivalent for experiment 2. The number of true positives decreases with increasing number of absence locations. This is reasonable taking into account the reduced area mapped as susceptible. Comparing the rating using the SSLI and the results of the assessment of the validation dataset shows that while the rating using the SSLI



**Fig. 8** **a** Location of the landslides recorded in the SSLI database (red dots), **b** Location of the SSLI landslides (black dots) on top of the reference shallow landslide susceptibility map for the canton Grisons. Purple areas indicate susceptibility to shallow landslide occurrence, **c** Shallow landslide susceptibility map for the canton Grisons (equal to Fig. 4). Purple areas indicate susceptibility to shallow landslide occurrence. Basemaps derived from the DHM25 (Federal Office of Topography swisstopo 2005)

shows the above described reduced accuracy with increasing number of absence locations, the validation dataset indicates for all strategies and ratios that the model is of good quality.

Table 5 shows the validation result of experiment 3. The validation results for all subsets show a reduced accuracy compared to the reference map. As with experiments 1 and 2, it can be seen that the validation using the validation dataset indicates models of higher quality than the validation using the SSLI which might be attributed to the clustered nature of the *Hangmuren* database.

The comparison between the results of the validation using the two different datasets and the thereby observed in parts strong discrepancy for all three experiments shows the added value of taking into account the independent SSLI as well.

Furthermore, this validation highlights that a larger variety of certain environmental conditions of the entries in the *Hangmuren* database would be desirable for the scale of the area of interest in the test case. However, it also underlines the importance of validation in general and the significance it should have in publications to reveal shortcomings and limitations.

### 3.7 Feature importance

The RF, as an inherently interpretable algorithm, provides feature importance information along with model training (Genier et al. 2010), giving an overview of the importance of the individual features for the RF prediction. This is done by comparing the results of the RF if the individual features would be removed from the input dataset (Taalab et al. 2018). Feature importance information can then be assessed using domain knowledge to identify model flaws through missing scientific consistency.

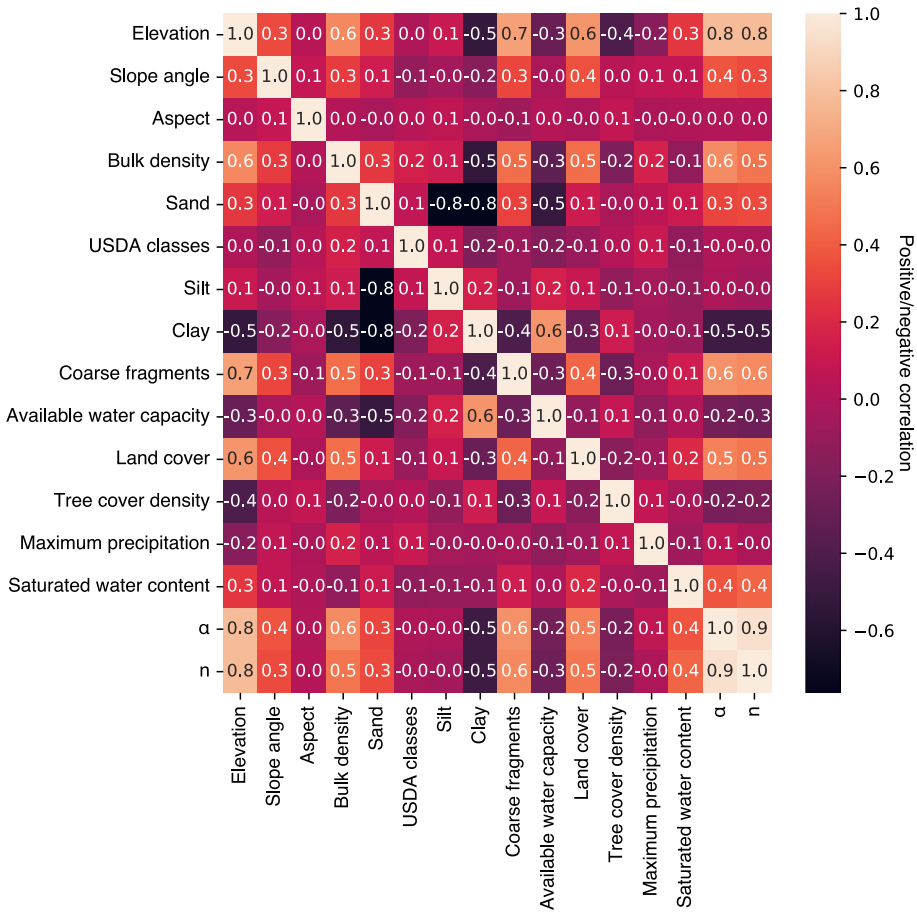
Figure 9 shows a correlation matrix of the features used in the test case. A certain degree of co-dependence is to be expected as earth is a complex system and different environmental domains are connected and cannot be regarded as isolated. A straightforward example for a correlation in the training dataset are the sand, silt and clay features that show the topsoil contents in percentage and sum up to 100%.

Several instances show strong correlation. Therefore, the RF feature importance output has to be assumed not to reflect the true importance. A manual approach to determining feature importance was adopted to resolve this issue.

The model was retrained manually after removing one feature or a group of features from the training dataset. The resulting map based on the retrained model was compared to the reference map to assess the influence of this feature or group of features. The feature groups were defined based on the correlation values but also taking domain knowledge into consideration. Some features showing stronger correlation like elevation and tree cover density should nevertheless not be grouped together as they come from different domains that physically influence the occurrence of landslides individually. Three feature groups were defined:

- *Group 1* bulk density, coarse fragments, elevation and land cover
- *Group 2*  $\alpha$  and  $n$
- *Group 3* sand, silt and clay content

Figure 10 shows the results of the manual feature importance evaluation. The ranking of the features from most important on the top to the least important on the bottom was done according to the percentage of identically predicted pixels with respect to the

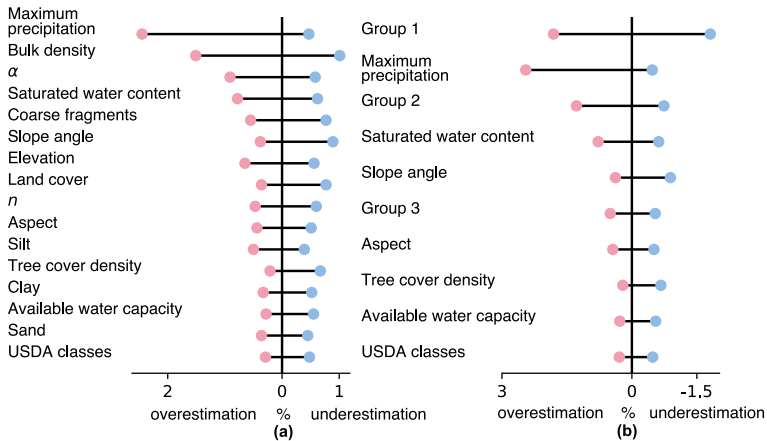


**Fig. 9** Correlation matrix of all features used in this study. Values close to 1 or - 1 respectively indicate a strong positive or negative correlation of the features

reference map. The larger the discrepancy, that is the smaller the percentage of identically predicted pixels, the more relevant is the feature for the mapping result.

Maximum observed precipitation and bulk density are the most important parameters. This is a sensible finding, due to the known importance of precipitation as a triggering and predisposing factor and bulk density as a factor controlling the infiltration of water into the ground. USDA classes, available water capacity, tree cover density as well as sand, silt and clay content are less important.

Overall, two conclusions can be drawn from the feature importance assessment. Firstly, a comparison of the original RF feature importance output and the manually derived ranking shows distinct discrepancies, justifying the adoption of an alternative approach to the standard RF feature importance output. Secondly, the feature importance ranking is scientifically consistent. This further strengthens the trust in the model. Generally, the importance of the individual features is quite similar as also the resulting susceptibility maps when removing the features are similar.



**Fig. 10** Feature importance according to the manual investigation. This figure illustrates the importance of individual features, represented by the percentage of pixels with different predictions compared to the reference model when each feature or feature group is removed from the training process. In blue, the percentage of pixels is indicated where the prediction shifted to 'not susceptible', suggesting an underestimation compared to the reference. In red, the percentage of pixels shows predictions of 'susceptible' where the reference found no susceptibility, indicating an overestimation. The features are ordered from top (most important) to bottom (least important) to visually convey their relative importance, **a** Feature importance ranking of all individual features, **b** Feature importance ranking of individual and groups of features. The feature group 1 comprises the features bulk density, coarse fragments, elevation and land cover. Group 2 comprises  $\alpha$  and  $n$  of the SWRC. Group 3 combines sand, silt and clay content

## 4 Discussion

We introduce a Python-based framework for susceptibility and hazard mapping, designed to facilitate future applications of RF-based map generation that is independent of secondary software. It allows the user to create reproducible maps in a user-friendly way through a generic implementation that is flexible in terms of area of interest, resolution and data basis. Therefore, the framework contains modular, scalable and transparent pre-implemented solutions for input dataset generation and mapping. The framework was successfully applied to a test case testing and demonstrating its reproducibility, extensibility and explainability. Three computational experiments were conducted using the framework to investigate the influence of the training dataset on the mapping result. These computational experiments are aimed to (1) explore sensitivities and limitations of the underlying RF method, and (2) support trust in the reliability of the framework through comparison with previous studies.

Experiment 1 involved a visual and qualitative evaluation of the impact of various feature combinations on the mapping result. The generated maps were assessed in comparison to the reference map. As expected, a single-sided feature composition resulted in maps with significant variations in the distribution and size of the susceptible areas. While there is no universally accepted strategy for selecting features [as highlighted by Reichenbach et al. (2018)], a common consensus is to strive for a balance of geospatial information from a range of environmental domains. However, the maps produced by models trained with subsets 1 and 2, both comprising a set of features from different environmental domains, still exhibit notable differences. While the general localisation of susceptible areas is

comparable, the total size of the susceptible area for subset 1 is significantly larger than for subset 2. Subset 2 also shows susceptibility in locations not captured by subset 1. This underscores the impact that the choice and combination of features have on the mapping result.

Studies investigating feature selection methods claim that the quality of mapping results can be increased by dropping irrelevant features (e.g. Pham et al. 2021; Nirbhav et al. 2023). Nirbhav et al. (2023) found that the set of chosen features depends strongly on the feature selection method. This is supported by the findings in Liu et al. (2021a) that feature selection is problem specific also with regard to the applied ML algorithm. While Nirbhav et al. (2023) describe that feature selection methods might decrease the accuracy of the resulting model, Kuhn and Johnson (2019) showed for the RF a decrease in accuracy for only a large amount of added irrelevant features. The number for which this decrease was observed exceeds by far the number of features typically included for ML-based landslide susceptibility and hazard mapping. Kumar et al. (2023) found that RF profited from an increased complexity by a higher number of features in comparison to other models tested in their study. Subsets 7 and 8 of this study (see Table 3) were created based on the feature importance assessment (see Fig. 10), and therefore reflect what the RF deems most important and least important. The resulting map when using only the most important features shows only small discrepancies to the reference map. The results therefore support the validity of the application of feature selection approaches. The difference in the maps using most and least important features also highlights the need to choose meaningful and representative features. While feature selection methods are regularly applied in landslide susceptibility and hazard mapping studies (e.g. Ado et al. 2022), considerations about the influence the chosen combination of features has on the resulting map independent of their individual importance are rarely considered in the literature even though its significance is shown here. Therefore, we recommend that upcoming studies on landslide susceptibility and hazard mapping using RF-based approaches take into account the feature combination as a factor of comparable significance to feature selection.

Experiment 2 investigated the influence of the sampling strategies of absence locations on the mapping result. Two separate investigations were carried out: (1) exploring the impact of the size and extent of the sampling area, and (2) examining the influence of the ratio of the number of presence to absence locations. Even though absence locations sampling has rarely been of interest in the past, several studies showed the importance of one or both parameters (Hong et al. 2019; Shao et al. 2020; Zhou et al. 2021; Wang et al. 2022). Table 4 and Fig. 6 show that the size of the absence location sampling area has a small influence on the area marked as susceptible in comparison to the effect of sampling ratio. The trend towards a reduced size of susceptible area for larger absence locations sampling areas as observed by Hong et al. (2019) and Shao et al. (2020) for coseismic landslides could not be reproduced. Wang et al. (2022) found a significant influence of the size of the sampling area for the quality of their resulting logistic regression model with larger sampling areas resulting in superior models. Overall, the results support the assumption that if absence locations are sampled in a representative way, they do not have to be sampled within the area of interest.

The ratio of the number of presence to absence locations significantly affects the mapping result. An increase in number of absence locations compared to presence locations results in a strong decrease of the size of the area mapped as susceptible. This trend matches the findings by Hong et al. (2019) and Shao et al. (2020). Hong et al. (2019) conclude that an equal ratio of presence to absence locations is to be preferred. Zhou et al. (2021) in contrast found that out of the ratios they tested 1:5 is most suitable

for their application case. Therefore, the choice of ratio is application-specific. As demonstrated by all examples, its influence on the result is important to consider during the conceptualisation of a research study and the interpretation of the results.

Finally, experiment 3 was conducted to investigate the importance of the representativeness of entries in the training dataset. The results lead to the conclusion that regional susceptibility mapping based on local presence data should be avoided. The literature review conducted for the present study highlights that discussions on the representativeness of data for the entire study area are often lacking. The number of landslides considered for training, relative to the size of the area of interest, varies significantly [e.g. 79 landslides in 49.74 km<sup>2</sup> (Hong et al. 2019), 841 in 2765 km<sup>2</sup> (Feng and Guo 2023), 132 in 33.4 km<sup>2</sup> (Vasu et al. 2016)]. The result of this experiment serves as a reminder to consider this influence when conceptualising a study and interpreting a map. The scarcity of landslide inventories in many areas often limits the options for optimising the data basis. The ML-based prediction is only as good as the data in terms of coverage, quality and representativeness. This should be taken into consideration in future studies and might as well guide future data acquisition efforts. Meyer and Pebesma (2021) suggest to establish an area of applicability for spatial prediction models to assess at which locations within the area of interest models can be reliably applied. This approach has, for instance, been employed by Betancourt et al. (2022) for global ozone predictions. We propose adopting this concept or a similar one for landslide susceptibility and hazard mapping to enhance the transparency of presented models and maps.

All of these results underscore, on the one hand, the necessity for careful consideration of mapping design choices, such as the composition of the training dataset. On the other hand, they affirm the validity and reliability of the results derived by applying the framework presented in this study.

## 5 Conclusions

We introduce a Python-based framework adhering to the *FAIR* principles, designed for the generic and flexible generation of landslide susceptibility and hazard maps. We show its application to a test case, demonstrating its reproducibility, extensibility and explainability and assess the results for their plausibility.

The produced maps are reasonable and the variations from the reference map align with expectations and findings in the literature. This provides confidence in the framework and in its resulting products which is a key prerequisite for its usage in future studies.

The presented results of the computational experiments complement existing knowledge, particularly regarding aspects that have been neglected in previous studies, such as absence locations sampling and training data representativeness.

Based on our observations, we recommend that future studies should place a stronger emphasis on: (1) the discussion and justification of the sampling strategy concerning absence locations, (2) the representativeness of the training data with respect to the area of interest, e.g. by applying the approach suggested by Meyer and Pebesma (2021) and (3) considering the influence of the combination of features just the way it is often done for the selection of features. These recommendations aim to enhance the robustness and reliability of the training dataset, ensuring a more effective and accurate outcome.



**Acknowledgements** We thank Dr. Ross Stirling from Newcastle University for his valuable comments which improved the manuscript. This work was performed as part of the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE).

**Author contributions** Ann-Kathrin Edrich: conceptualisation, methodology, software, validation, formal analysis, investigation, writing—original draft, writing—review and editing, visualization; Anil Yildiz: conceptualisation, methodology, writing—review and editing, supervision; Ribana Roscher: methodology, writing—review and editing, funding acquisition; Alexander Bast: formal analysis, writing—review and editing; Frank Graf: formal analysis, writing—review and editing; Julia Kowalski: conceptualisation, methodology, resources, writing—review and editing, supervision, project administration, funding acquisition

**Funding** Open Access funding enabled and organized by Projekt DEAL. Funding has been provided through the German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection under Grant No 67KI2043 (KISTE project).

**Data availability** All datasets used in this study with the exception of the SSLI are publicly-available. For more information on the SSLI please contact Alexander Bast (WSL-SLF, alexander.bast@slf.ch). The landslide susceptibility and hazard mapping framework presented in this study can be accessed via Edrich et al. (2023).

## Declarations

**Conflict of interest** The authors have no conflict of interest to declare that are relevant to the content of this article.

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

## References

- Abraham MT, Satyam N, Lokesh R, Pradhan B, Alamri A (2021) Factors affecting landslide susceptibility mapping: assessing the influence of different machine learning approaches, sampling strategies and data splitting. *Land* 10(9):989. <https://doi.org/10.3390/land10090989>
- Ado M, Amitab K, Maji AK, Jasińska E, Gono R, Leonowicz Z, Jasiński M (2022) Landslide susceptibility mapping using machine learning: a literature survey. *Remote Sens* 14(13):3029. <https://doi.org/10.3390/rs14133029>
- Arora M, Das Gupta A, Gupta R (2004) An artificial neural network approach for landslide hazard zonation in the Bhagirathi (Ganga) valley, Himalayas. *Int J Remote Sens* 25(3):559–572. <https://doi.org/10.1080/0143116031000156819>
- Ballabio C, Panagos P, Monatanarella L (2016) Mapping topsoil physical properties at European scale using the LUCAS database. *Geoderma* 261:110–123. <https://doi.org/10.1016/j.geoderma.2015.07.006>
- Bast A, Wilcke W, Graf F, Lüscher P, Gärtner H (2016) Does mycorrhizal inoculation improve plant survival, aggregate stability, and fine root development on a coarse-grained soil in an alpine eco-engineering field experiment? *J Geophys Res Biogeosci* 121(8):2158–2171. <https://doi.org/10.1002/2016JG003422>
- Baum RL, Savage WZ, Godt JW (2002) TRIGRS—a fortran program for transient rainfall infiltration and grid-based regional slope-stability analysis. *US Geol Surv Open-File Rep* 424:38
- Bebi P, Bast A, Ginzler C, Rickli C, Schöngrundner K, Graf F (2019) Waldentwicklung und flachgründige rutschungen: Eine grossflächige gis-analyse. *Schweiz Z für Forstwes* 170(6):318–325. <https://doi.org/10.3188/szf.2019.0318>
- Belgiu M, Drăguț L (2016) Random forest in remote sensing: a review of applications and future directions. *ISPRS J Photogramm Remote Sens* 114:24–31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>

- Betancourt C, Stomberg TT, Edrich AK, Patnala A, Schultz MG, Roscher R, Kowalski J, Stadler S (2022) Global, high-resolution mapping of tropospheric ozone-explainable machine learning and impact of uncertainties. *Geosci Model Dev* 15(11):4331–4354. <https://doi.org/10.5194/gmd-15-4331-2022>
- Bragagnolo L, da Silva RV, Grzybowski JMV (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 (2001) Random forests. *Mach Learn* 45:5–32. <https://doi.org/10.1023/A:1010933404324>
- Bui DT, Pradhan B, Lofman O, Revhaug I, Dick OB (2012) Landslide susceptibility assessment in the Hoa Binh province of Vietnam: a comparison of the Levenberg–Marquardt and Bayesian regularized neural networks. *Geomorphology* 171:12–29. <https://doi.org/10.1016/j.geomorph.2012.04.023>
- Caine N (1980) The rainfall intensity-duration control of shallow landslides and debris flows. *Geogr Ann Ser A Phys Geogr* 62(1–2):23–27. <https://doi.org/10.1080/04353676.1980.11879996>
- Chen W, Xie X, Wang J, Pradhan B, Hong H, Bui DT, Duan Z, Ma J (2017) A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena* 151:147–160. <https://doi.org/10.1016/j.catena.2016.11.032>
- Conforti M, Pascale S, Robustelli G, Sdao F (2014) Evaluation of prediction capability of the artificial neural networks for mapping landslide susceptibility in the Turbolo River catchment (Northern Calabria, Italy). *Catena* 113:236–250. <https://doi.org/10.1016/j.catena.2013.08.006>
- Copernicus Land Monitoring Service (2018a) Corine land cover (CLC) 2018, version 2020\_20u1. <https://land.copernicus.eu/pan-european/corine-land-cover/clc2018?tab=metadata>. Accessed 09 2021
- Copernicus Land Monitoring Service (2018b) High resolution layer: tree cover density (TCD) 2018. <https://land.copernicus.eu/pan-european/high-resolution-layers/forests/tree-cover-density/status-maps/tree-cover-density-2018?tab=metadata>. Accessed 09 2021
- Cutler A, Cutler DR, Stevens JR (2012) Random forests. In: Zhang C, Ma Y (eds) *Ensemble machine learning: methods and applications*. Springer, New York, pp 157–175. [https://doi.org/10.1007/978-1-4419-9326-7\\_5](https://doi.org/10.1007/978-1-4419-9326-7_5)
- Dang VH, Dieu TB, Tran XL, Hoang ND (2019) Enhancing the accuracy of rainfall-induced landslide prediction along mountain roads with a GIS-based random forest classifier. *Bull Eng Geol Environ* 78(4):2835–2849. <https://doi.org/10.1007/s10064-018-1273-y>
- Dong A, Dou J, Fu Y, Zhang R, Xing K (2023) Unraveling the evolution of landslide susceptibility: a systematic review of 30-years of strategic themes and trends. *Geocarto Int* 38(1):2256308. <https://doi.org/10.1080/10106049.2023.2256308>
- Dou J, Yamagishi H, Pourghasemi HR, Yunus AP, Song X, Xu Y, Zhu Z (2015) An integrated artificial neural network model for the landslide susceptibility assessment of Osado Island, Japan. *Nat Hazards* 78(3):1749–1776. <https://doi.org/10.1007/s11069-015-1799-2>
- Dou J, Yunus AP et al (2019) Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. *Sci Total Environ* 662:332–346. <https://doi.org/10.1016/j.scitotenv.2019.01.221>
- Du J, Glade T, Woldai T, Chai B, Zeng B (2020) Landslide susceptibility assessment based on an incomplete landslide inventory in the Jilong valley, Tibet, Chinese Himalayas. *Eng Geol* 270:105572. <https://doi.org/10.1016/j.enggeo.2020.105572>
- Edrich AK, Yildiz A, Kowalski J (2023) Landslide susceptibility and hazard mapping framework. <https://doi.org/10.6084/m9.figshare.24339643>
- Environmental Systems Research Institute, Inc. (2010a) ArcGIS. <https://www.esri.com/en-us/arcgis>. Accessed 12 2023
- Environmental Systems Research Institute, Inc. (2010b) ArcGIS Pro. <https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview>. Accessed 12 2023
- Federal Office of Topography Swisstopo (2005) DHM25. <https://www.swisstopo.admin.ch/en/geodata/height/dhm25.html>. Accessed 09 2021
- Feng L, Guo M et al (2023) Comparative analysis of machine learning methods and a physical model for shallow landslide risk modeling. *Sustainability* 15(1):6. <https://doi.org/10.3390/su15010006>
- Gaidzik K, Ramírez-Herrera MT (2021) The importance of input data on landslide susceptibility mapping. *Sci Rep* 11(1):19334. <https://doi.org/10.1038/s41598-021-98830-y>
- Genuer R, Poggi JM, Tuleau-Malot C (2010) Variable selection using random forests. *Pattern Recognit Lett* 31(14):2225–2236. <https://doi.org/10.1016/j.patrec.2010.03.014>
- Graf F, Bast A, Gärtner H, Yildiz A (2019) Effects of mycorrhizal fungi on slope stabilisation functions of plants. In: Wu W (ed) *Recent advances in geotechnical research*. Springer, Cham, pp 57–77. [https://doi.org/10.1007/978-3-319-89671-7\\_6](https://doi.org/10.1007/978-3-319-89671-7_6)

- Hao J, Ho TK (2019) Machine learning made easy: a review of Scikit-learn package in python programming language. *J Educ Behav Stat* 44(3):348–361. <https://doi.org/10.3102/10769986198322>
- Hastie T, Tibshirani R, Friedman J (2009) Random forests. In: Hastie T, Tibshirani R, Friedman J (eds) *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York, pp 587–604. [https://doi.org/10.1007/978-0-387-84858-7\\_15](https://doi.org/10.1007/978-0-387-84858-7_15)
- Hervás J, Bobrowsky P (2009) Mapping: inventories, susceptibility, hazard and risk. In: Sassa K, Canuti P (eds) *Landslides-disaster risk reduction*. Springer, Berlin, pp 321–349. [https://doi.org/10.1007/978-3-540-69970-5\\_19](https://doi.org/10.1007/978-3-540-69970-5_19)
- Hong H, Miao Y, Liu J, Zhu AX (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>
- Huang W, Ding M, Li Z, Zhuang J, Yang J, Li X, Meng L, Zhang H, Dong Y (2022) An efficient user-friendly integration tool for landslide susceptibility mapping based on support vector machines: SVM-LSM toolbox. *Remote Sens* 14(14):3408. <https://doi.org/10.3390/rs14143408>
- Huffman G, Stocker E, Bolvin D, Nelkin E, Tan J (2019) GPM IMERG final precipitation L3 1 month 0.1 degree x 0.1 degree V06. <https://doi.org/10.5067/GPM/IMERG/3B-MONTH/06>. Accessed 09 2021
- Karantanellis E, Marinos V, Vassilakis E, Hölbling D (2021) Evaluation of machine learning algorithms for object-based mapping of landslide zones using UAV data. *Geosciences* 11(8):305. <https://doi.org/10.3390/geosciences11080305>
- Kavzoglu T, Colkesen I, Sahin EK (2019) Machine learning techniques in landslide susceptibility mapping: A survey and a case study. In: Pradhan SP, Vishal V, Singh TN (eds) *Landslides: theory, practice and modelling*. Springer, Cham, pp 283–301. [https://doi.org/10.1007/978-3-319-77377-3\\_13](https://doi.org/10.1007/978-3-319-77377-3_13)
- Kinkeldey C, MacEachren AM, Riveiro M, Schiewe J (2017) Evaluating the effect of visually represented geodata uncertainty on decision-making: systematic review, lessons learned, and recommendations. *Cartogr Geogr Inf Sci* 44(1):1–21. <https://doi.org/10.1080/15230406.2015.1089792>
- Kuhn M, Johnson K (2019) Feature engineering and selection: a practical approach for predictive models. Taylor & Francis, Boca Raton. <https://doi.org/10.1201/9781315108230>
- Kumar C, Walton G, Santi P, Luza C (2023) An ensemble approach of feature selection and machine learning models for regional landslide susceptibility mapping in the arid mountainous terrain of southern Peru. *Remote Sens* 15(5):1376. <https://doi.org/10.3390/rs15051376>
- Kumar D, Thakur M, Dubey CS, Shukla DP (2017) Landslide susceptibility mapping and prediction using support vector machine for Mandakini River basin, Garhwal Himalaya, India. *Geomorphology* 295:115–125. <https://doi.org/10.1016/j.geomorph.2017.06.013>
- Kuradusenge M, Kumaran S, Zennaro M (2020) Rainfall-induced landslide prediction using machine learning models: the case of Ngororero district, Rwanda. *Int J Environ Res Public Health* 17(11):4147. <https://doi.org/10.3390/ijerph17114147>
- Lamprecht AL, Garcia L et al (2020) Towards FAIR principles for research software. *Data Sci* 3(1):37–59. <https://doi.org/10.3233/DS-190026>
- Lee S, Choi J, Woo I (2004) The effect of spatial resolution on the accuracy of landslide susceptibility mapping: a case study in Boun, Korea. *Geosci J* 8:51–60. <https://doi.org/10.1007/BF02910278>
- Leonarduzzi E, Molnar P, McArdell BW (2017) Predictive performance of rainfall thresholds for shallow landslides in Switzerland from gridded daily data. *Water Resour Res* 53(8):6612–6625. <https://doi.org/10.1002/2017WR021044>
- Lima P, Steger S, Glade T, Murillo-García FG (2022) Literature review and bibliometric analysis on data-driven assessment of landslide susceptibility. *J Mt Sci* 19(6):1670–1698. <https://doi.org/10.1007/s11629-021-7254-9>
- Liu LL, Yang C, Wang XM (2021a) Landslide susceptibility assessment using feature selection-based machine learning models. *Geomech Eng* 25(1):1–16
- Liu Z, Gilbert G, Cepeda JM, Lysdahl AOK, Piciullo L, Hefre H, Lacasse S (2021b) Modelling of shallow landslides with machine learning algorithms. *Geosci Front* 12(1):385–393. <https://doi.org/10.1016/j.gsf.2020.04.014>
- Liu W, Zhang Y, Liang Y, Sun P, Li Y, Su X, Wang A, Meng X (2022) Landslide risk assessment using a combined approach based on InSAR and random forest. *Remote Sens* 14(9):2131. <https://doi.org/10.3390/rs14092131>
- 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>
- Maleki M, Mir Mohammad Hosseini SM (2022) Assessment of the pseudo-static seismic behavior in the soil nail walls using numerical analysis. *Innov Infrastruct Solut* 7(4):262. <https://doi.org/10.1007/s41062-022-00861-5>

- Meyer H, Pebesma E (2021) Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods Ecol Evol* 12(9):1620–1633. <https://doi.org/10.1111/2041-210X.13650>
- Neteler M, Bowman MH, Landa M, Metz M (2012) GRASS GIS: a multi-purpose open source GIS. *Environ Modell Softw* 31:124–130. <https://doi.org/10.1016/j.envsoft.2011.11.014>
- Nirbhav Malik A, Maheshwar Jan T, Prasad M (2023) Landslide susceptibility prediction based on decision tree and feature selection methods. *J Indian Soc Remote Sens* 51:771–786. <https://doi.org/10.1007/s12524-022-01645-1>
- Nurwati N, Ummah MH, Cahyono AB, Darminto MR, Hong JH (2022) A comparison study of landslide susceptibility spatial modeling using machine learning. *ISPRS Int J Geo-Inf* 11(12):602. <https://doi.org/10.3390/ijgi11120602>
- Osna T, Sezer EA, Akgun A (2014) GeoFIS: an integrated tool for the assessment of landslide susceptibility. *Comput Geosci* 66:20–30. <https://doi.org/10.1016/j.cageo.2013.12.016>
- Pandey VK, Sharma MC (2017) Probabilistic landslide susceptibility mapping along Tipri to Ghuttu highway corridor, Garhwal Himalaya (India). *Remote Sens Appl Soc Environ* 8:1–11. <https://doi.org/10.1016/j.rsase.2017.07.007>
- Pedregosa F, Varoquaux G et al (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830
- Pham BT, Van Dao D, Acharya TD, Van Phong T, Costache R, Van Le H, Nguyen HBT, Prakash I (2021) Performance assessment of artificial neural network using chi-square and backward elimination feature selection methods for landslide susceptibility analysis. *Environ Earth Sci* 80:1–13. <https://doi.org/10.1007/s12665-021-09998-5>
- Pradhan AMS, Kim YT (2020) Rainfall-induced shallow landslide susceptibility mapping at two adjacent catchments using advanced machine learning algorithms. *ISPRS Int J Geo-Inf* 9(10):569. <https://doi.org/10.3390/ijgi9100569>
- Python Software Foundation (2021) Python programming language. <https://www.python.org/>. Accessed 12 2023
- QGIS Development Team (2020) QGIS geographic information system. <https://qgis.org/>. Accessed 12 2023
- R Core Team (2020) R: a language and environment for statistical computing. <https://www.R-project.org/>. Accessed 12 2023
- Regmi NR, Giardino JR, McDonald EV, Vitek JD (2014) A comparison of logistic regression-based models of susceptibility to landslides in western Colorado, USA. *Landslides* 11(2):247–262. <https://doi.org/10.1007/s10346-012-0380-2>
- Reichenbach P, Rossi M, Malamud BD, Mihir M, Guzzetti F (2018) A review of statistically-based landslide susceptibility models. *Earth-Sci Rev* 180:60–91. <https://doi.org/10.1016/j.earscirev.2018.03.001>
- Rickli C, Graf F, Bebi P, Bast A, Loup B, McArdell B (2019) Schützt der wald vor rutschungen? Hinweise aus der wsl-rutschungsdatenbank. *Schweiz Z für Forstwes* 170(6):310–317. <https://doi.org/10.3188/szf.2019.0310>
- Saha S, Roy J, Pradhan B, Hembram TK (2021) Hybrid ensemble machine learning approaches for landslide susceptibility mapping using different sampling ratios at East Sikkim Himalayan, India. *Adv Space Res* 68(7):2819–2840. <https://doi.org/10.1016/j.asr.2021.05.018>
- Sahin EK, Colkesen I, Acemali SS, Akgun A, Aydinoglu AC (2020) Developing comprehensive geocomputation tools for landslide susceptibility mapping: LSM tool pack. *Comput Geosci* 144:104592. <https://doi.org/10.1016/j.cageo.2020.104592>
- Sevgen E, Kocaman S, Nefeslioglu HA, Gokceoglu C (2019) A novel performance assessment approach using photogrammetric techniques for landslide susceptibility mapping with logistic regression, ANN and random forest. *Sensors* 19(18):3940. <https://doi.org/10.3390/s19183940>
- Sezer EA, Nefeslioglu HA, Osna T (2017) An expert-based landslide susceptibility mapping (LSM) module developed for Netcad Architect Software. *Comput Geosci* 98:26–37. <https://doi.org/10.1016/j.cageo.2016.10.001>
- Shano L, Raghuvanshi TK, Meten M (2020) Landslide susceptibility evaluation and hazard zonation techniques—a review. *Geoenviron Disasters* 7(1):1–19. <https://doi.org/10.1186/s40677-020-00152-0>
- Shao X, Ma S, Xu C, Zhou Q (2020) Effects of sampling intensity and non-slide/slide sample ratio on the occurrence probability of coseismic landslides. *Geomorphology* 363:107222. <https://doi.org/10.1016/j.geomorph.2020.107222>
- Shirzadi A, Soliamani K et al (2018) Novel GIS based machine learning algorithms for shallow landslide susceptibility mapping. *Sensors* 18(11):3777. <https://doi.org/10.3390/s18113777>
- Stanley TA, Kirschbaum DB, Benz G, Emberson RA, Amatya PM, Medwedeff W, Clark MK (2021) Data-driven landslide nowcasting at the global scale. *Front Earth Sci* 9:378. <https://doi.org/10.3389/feart.2021.640043>

- Stumpf A, Kerle N (2011) Object-oriented mapping of landslides using random forests. *Remote Sens Environ* 115(10):2564–2577. <https://doi.org/10.1016/j.rse.2011.05.013>
- Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Research Unit Mountain Hydrology and Mass Movements (2023a) Datenquelle hangmuren-datenbank. <https://hangmuren.wsl.ch/>. Accessed 03 2021
- Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Research Unit Mountain Hydrology and Mass Movements (2023b) Supplemented swiss landslide inventory. Access upon request (alexander.bast@slf.ch)
- Taalab K, Cheng T, Zhang Y (2018) Mapping landslide susceptibility and types using random forest. *Big Earth Data* 2(2):159–178. <https://doi.org/10.1080/20964471.2018.1472392>
- Thiery Y, Terrier M, Colas B, Fressard M, Maquaire O, Grandjean G, Gourdiere S (2020) Improvement of landslide hazard assessments for regulatory zoning in France: state-of-the-art perspectives and considerations. *Int J Disaster Risk Reduct* 47:101562. <https://doi.org/10.1016/j.ijdrr.2020.101562>
- Tien Bui D, Tuan TA, Klempe H, Pradhan B, Revhaug I (2016) Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides* 13(2):361–378. <https://doi.org/10.1007/s10346-015-0557-6>
- Tóth B, Weynants M, Pásztor L, Hengl T (2017) 3D soil hydraulic database of Europe at 250 m resolution. *Hydrol Process* 31(14):2662–2666. <https://doi.org/10.1002/hyp.11203>
- Trigila A, Iadanza C, Esposito C, Scarascia-Mugnozza G (2015) Comparison of logistic regression and random forests techniques for shallow landslide susceptibility assessment in Giampilieri (NE Sicily, Italy). *Geomorphology* 249:119–136. <https://doi.org/10.1016/j.geomorph.2015.06.001>
- Uchida T, Osanai N, Onoda S, Takayama T, Tomura K (2006) A simple method for producing probabilistic seismic shallow landslide hazard maps. In: *Proceedings of interpraevent*, pp 529–534
- Van Den Eeckhaut M, Reichenbach P, Guzzetti F, Rossi M, Poesen J (2009) Combined landslide inventory and susceptibility assessment based on different mapping units: an example from the Flemish Ardennes, Belgium. *Nat Hazards Earth Syst Sci* 9(2):507–521. <https://doi.org/10.5194/nhess-9-507-2009>
- Vasu NN, Lee SR, Pradhan AMS, Kim YT, Kang SH, Lee DH (2016) A new approach to temporal modelling for landslide hazard assessment using an extreme rainfall induced-landslide index. *Eng Geol* 215:36–49. <https://doi.org/10.1016/j.enggeo.2016.10.006>
- Wang Z, Liu Q, Liu Y (2020) Mapping landslide susceptibility using machine learning algorithms and GIS: A case study in Shexian county, Anhui province, China. *Symmetry* 12(12):1954. <https://doi.org/10.3390/sym12121954>
- Wang C, Lin Q, Wang L, Jiang T, Su B, Wang Y, Mondal SK, Huang J, Wang Y (2022) The influences of the spatial extent selection for non-landslide samples on statistical-based landslide susceptibility modelling: a case study of Anhui province in China. *Nat Hazards* 112(3):1967–1988. <https://doi.org/10.1007/s11069-022-05252-8>
- Wilkinson MD, Dumontier M et al (2016) The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 3(1):1–9. <https://doi.org/10.1038/sdata.2016.18>
- Xiao C, Tian Y, Shi W, Guo Q, Wu L (2010) A new method of pseudo absence data generation in landslide susceptibility mapping with a case study of Shenzhen. *Sci China Technol Sci* 53(1):75–84. <https://doi.org/10.1007/s11431-010-3219-x>
- Yeon YK, Han JG, Ryu KH (2010) Landslide susceptibility mapping in Injae, Korea, using a decision tree. *Eng Geol* 116(3–4):274–283. <https://doi.org/10.1016/j.enggeo.2010.09.009>
- Youssef AM, Pourghasemi HR (2021) Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha basin, Asir region, Saudi Arabia. *Geosci Front* 12(2):639–655. <https://doi.org/10.1016/j.gsf.2020.05.010>
- Zhang Y, Wu W et al (2020) Mapping landslide hazard risk using random forest algorithm in Guixi, Jiangxi, China. *ISPRS Int J Geo-Inf* 9(11):695. <https://doi.org/10.3390/ijgi9110695>
- Zhou X, Wen H, Zhang Y, Xu J, Zhang W (2021) Landslide susceptibility mapping using hybrid random forest with geodetector and RFE for factor optimization. *Geosci Front* 12(5):101211. <https://doi.org/10.1016/j.gsf.2021.101211>
- Zhu AX, Miao Y, Liu J, Bai S, Zeng C, Ma T, Hong H (2019) A similarity-based approach to sampling absence data for landslide susceptibility mapping using data-driven methods. *Catena* 183:104188. <https://doi.org/10.1016/j.catena.2019.104188>

## Authors and Affiliations

**Ann-Kathrin Edrich**<sup>1</sup>  · **Anil Yildiz**<sup>1</sup>  · **Ribana Roscher**<sup>2,3</sup>  · **Alexander Bast**<sup>4,5</sup>  · **Frank Graf**<sup>5,6</sup>  · **Julia Kowalski**<sup>1</sup> 

✉ Ann-Kathrin Edrich  
edrich@mbd.rwth-aachen.de

- <sup>1</sup> Methods for Model-based Development in Computational Engineering, RWTH Aachen University, Aachen, Germany
- <sup>2</sup> Institute of Bio- and Geosciences, Plant Sciences (IBG-2), Forschungszentrum Jülich GmbH, Jülich, Germany
- <sup>3</sup> Institute of Geodesy and Geoinformation, Remote Sensing Group, University of Bonn, Bonn, Germany
- <sup>4</sup> Alpine Environment and Natural Hazards/Permafrost, WSL Institute for Snow and Avalanche Research SLF, Davos Dorf, Switzerland
- <sup>5</sup> Climate Change, Extremes and Natural Hazards in Alpine Regions Research Centre CERC, WSL Institute for Snow and Avalanche Research SLF, Davos Dorf, Switzerland
- <sup>6</sup> Alpine Environment and Natural Hazards/Mountain Ecosystems, WSL Institute for Snow and Avalanche Research SLF, Davos Dorf, Switzerland