Acta Geotechnica
https://doi.org/10.1007/s11440-024-02384-y

RESEARCH PAPER

Landslide susceptibility mapping using physics-guided machine learning: a case study of a debris flow event in Colorado Front Range

Te Pei[1] · Tong Qiu[2]

Received: 14 October 2023 / Accepted: 27 July 2024

Abstract
Landslides are common geohazards worldwide, resulting in significant losses to economies and human lives. Data-driven approaches, especially machine learning (ML) models, have been widely used recently for landslide susceptibility mapping (LSM) by extracting features from geospatial variables based on their contribution to landslide occurrences using known distributions of landslides as the training dataset. However, challenges remain in applying ML models for LSM models due to the scarcity and uneven spatial distribution of landslide data coupled with the spatial heterogeneity of hillslope conditions. Moreover, ML models developed with limited data often exhibit unexpected behaviors, resulting in poor interpretability and predictions that deviate from intuitive expectations and established domain knowledge. To overcome these challenges, this study proposes a physics-guided machine learning (PGML) framework that integrates landslide domain knowledge into ML models for LSM. The PGML framework was developed and assessed using a detailed debris flow inventory from a storm event in the Colorado Front Range. Based on the infinite slope model, the factor of safety for the study area was first determined and was subsequently used to constrain the prediction of ML models through a modified loss function and measure the physics consistency of model predictions. To evaluate the robustness and generalizability of the models, this study uses geographical sample selections for model performance evaluation, where ML models are trained and tested across heterogeneous ecoregions. The results of this study demonstrated the efficacy of both physics-based and data-driven methods in determining landslide susceptibility in the study area; however, pure data-driven ML models produced physically unrealistic results and poor generalization performance in new ecoregions. With the incorporation of physical constraints, the PGML model demonstrated notable enhancements in physics consistency and generalization capability, along with reduced model uncertainties across various ecoregions, surpassing the performance of benchmark ML models.

Keywords Case study · Debris flow · Landslide susceptibility mapping · Machine learning · Physics-based model

1 Introduction

Landslides represent a major natural hazard, causing substantial economic and human losses globally [22, 36]. In the USA, landslides occur in nearly all states and are a recurrent and significant issue, primarily in coastal and mountainous regions such as the Pacific Northwest, California, and the Appalachian Mountains. These events result in approximately 25–50 fatalities and exceed one billion dollars in losses annually [53]. In addition, more frequent extreme climate events, such as heavy rainfall and wildfires, are expected under current climate projections [21, 108], which may result in more frequent landslides and related damages. As urban expansion progresses into mountainous regions, many infrastructures are becoming susceptible to the threats posed by landslides [29]. Consequently, pinpointing potential landslide zones becomes pivotal in elevating public awareness and lessening foreseeable repercussions.

✉ Tong Qiu
tong.qiu@utah.edu

Te Pei
tpei@ccny.cuny.edu

[1] Department of Civil Engineering, The City University of New York (City College), New York, NY 10031, USA

[2] Department of Civil and Environmental Engineering, The University of Utah, Salt Lake City, UT 84112, USA

Published online: 13 August 2024

Landslide susceptibility mapping (LSM) is a systematic process that identifies and quantifies the susceptibility of landslide occurrences in specified geographic regions based on the analysis of various contributing factors such as geologic formations, topographic characteristics, and climatic conditions. It answers the question of where landslides are likely to occur [24]. As outlined by Reichenbach et al. [72] and Merghadi et al. [52], strategies to identify zones prone to landslides primarily fall into three categories: physics-based, heuristic, and statistical methods. The physics-based strategies depend heavily on the limit equilibrium analysis grounded in soil mechanics, which defines the limiting state where the shear stress along a potential failure surface in the slope has reached the shear strength of the slope materials. Notably, methodologies such as the infinite slope model, along with its variations, have become popular tools in assessing landslide risk for rainfall-induced shallow landslides (e.g., [7, 33, 49, 55, 57, 91]). The physics-based methods can provide physically consistent results that align with geotechnical domain knowledge on landslide mechanism; however, their implementation in LSM is often confined to small-scale analysis due to simplified physical assumptions embedded in physical models and the challenges of obtaining precise and comprehensive data on soil and hydrological attributes. Due to these limitations, physics-based methods for LSM are typically employed to provide early warning for impending slope failures [61, 90], where detailed site characterization can be obtained. Heuristic approaches use opinion-driven models that conduct landslide susceptibility zonation by ranking and weighting instability factors based on expert opinion and expertise (e.g., [12, 60, 70]). However, results based on heuristic approaches are challenging to evaluate and quantify objectively as they rely on investigators' understanding and judgments on the actual causes of landslides in the study area [72]. Statistically based approaches create functional correlations between landslide susceptibility and various geo-environmental determinants based on the analysis of historic and ongoing landslide location observations. The recent developments in remote sensing and data science have been steering the research focus toward these statistically based methodologies [72]. Methods for statistically based approaches include classic statistical analysis and machine learning (ML), and substantial progress has been made using various techniques and algorithms. Detailed reviews of statistical and ML methods for landslide susceptibility modeling and associated terrain zonations can be found in [43, 52, 72].

Although statistically based approaches, particularly ML models, have made tremendous progress in recent LSM studies, they encounter two inherent challenges that limit their effectiveness: data availability and model reliability. Moreover, it should be noted that these issues are not unique to LSM but are prevalent in general ML applications across various disciplines (e.g., [56, 58]). For the data availability challenges, data-driven LSM models heavily depend on data availability and quality, which presents significant obstacles for regional and broader-scale applications. Existing landslide inventories are often sparse and unevenly distributed (i.e., they have bias and insufficient representation), and substantial environmental variations exist between hillslopes and different regions [53, 95]. Landslide inventories that are consistent, accurate, representative, and cover extensive regions remain very limited [53]. Despite some recent advancements in automating landslide inventory mapping from remote sensing images (e.g., [59, 64, 106]), these challenges persist in effectively applying data-driven LSM models on a larger scale. The model reliability challenge is another critical but often overlooked issue. The inherent flexibility of ML models (e.g., deep neural networks or tree-based models) enables them to capture nuances in large datasets; however, this flexibility can also lead to unexpected behaviors in parts of the input space not covered by the training and validation datasets (e.g., [28, 67, 94, 104]), potentially failing to reflect the fundamental physics behind mass movements of landslides (e.g., [67, 81, 103]). This issue is especially problematic when learning from small datasets, which can lead to overfitting and poor generalization. Importantly, these problems remain undetected during the development phase due to dataset limitations [104].

Therefore, there is a critical need to develop robust models and comprehensive model evaluation frameworks that can navigate through these challenges, thereby ensuring the reliability and applicability of data-driven LSM models across regional or larger scales. In response to these challenges, recent progress in the community has included the adoption of ensemble methods [19] and knowledge transfer strategies such as transfer learning [105] and few-shot learning [92]. Ensemble methods help improve model stability and accuracy by aggregating predictions from multiple models, thereby reducing the risk of overfitting by improving statistical robustness in LSM applications (e.g., [20, 23, 42, 107]). Meanwhile, transfer learning has been utilized to effectively improve LSM model performance by transferring knowledge learned from source regions to target regions where data may be limited (e.g., [93, 96, 103]). Few-shot learning, on the other hand, enables rapid model adaptation to new, often scarcely sampled areas, thus addressing the data availability challenge [16]. Despite these advances, these methodologies are primarily focused on improving model performance based on accuracy metrics [78]; they often overlook the integration of domain-specific knowledge and physical laws, which is crucial for accurately modeling

susceptibility for complex geological processes like landslides. This oversight is significant because of the aforementioned challenges. Limited, non-representative data and unconstrained ML flexibility can lead to overfitting and models that poorly reflect actual landslide susceptibility during deployment. For example, Pei et al. [66] and Pei and Qiu [67] evaluated commonly used ML models for predicting slope stabilities using case histories for circular failure slope; their results showed that ML models with good performance based on data science accuracy metrics could behave poorly in terms of physics consistency, which highlights the need for innovative approaches that combine data-driven approaches with domain knowledge.

Physics-based models can estimate landslide susceptibility without the need for labeled data, relying solely on established domain knowledge to interpret geological conditions and predict potential landslide areas. However, data-driven approaches often overlook the insights offered by these physics-based models, leaving a significant gap in harnessing the potential of domain knowledge within these systems. Bridging this gap represents a substantial opportunity to enhance the accuracy and robustness of LSM models by integrating empirical data with theoretical understanding. Incorporating domain-specific knowledge into data-driven models has gained traction in various fields, including computer vision and natural language processing, and some regulated domains that require high-stakes predictions, such as healthcare, criminal justice, and finance, as detailed in comprehensive surveys by Carvalho et al. [15] and von Rueden et al. [88]. In the case of modeling physics processes, ML models that integrate physics principles are generally termed physics-guided machine learning (PGML), as mentioned in several studies (e.g., [30, 38, 45, 63, 66, 74]). PGML aims to use ML techniques to model physical phenomena while incorporating the underlying physical laws and constraints into the model. This approach can help improve the model performance and make predictions more physically meaningful. Figure 1 presents a schematic comparison of the uses of data and theory among different models. PGML has been applied in various fields, including geoscience

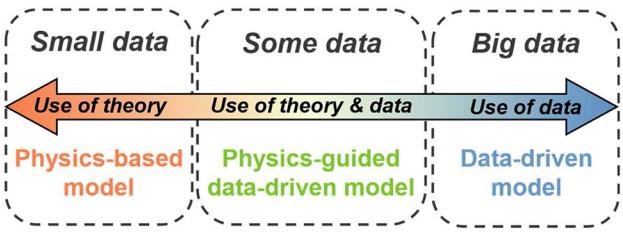

Fig. 1 Comparison of model development strategies across varying data scales (after Pei et al. [66])

applications, to improve the accuracy and interpretability of ML models and uncover new physical insights.

There are several different ways that physics can be incorporated into ML models. For example, hybrid physics and data-driven (HPD) models with data augmentations and feature enhancements based on physics-based models have been used to predict various scientific problems, and improvements in model performances have been observed in multiple studies (e.g., [18, 26, 99, 100, 102]). Apart from HPD models, physics-based models can also provide foundations for initializing and informing data-driven models [46, 66, 67]. For example, Read et al. [73] and Jia et al. [30] enhanced the accuracy and robustness of their deep learning (DL) models for lake temperature forecasting using pre-trained models based on simulated data from physics-based models. Moreover, Ma et al. [54] demonstrated the significance of knowledge transfer; they applied weights learned from a comprehensive USA dataset to regions with less data availability, such as China and Chile. In addition, physics-based model architectures can be used to constrain data-driven model behaviors. For example, Rahmani et al. [76] and Tsai et al. [85] implemented process-based models as differentiable modules into the deep learning framework for hydrological modeling, and performance improvement was observed on various metrics. Moreover, physics-based regularization can be used to impose physical constraints on the model to ensure the model behaves in a physically meaningful way. For instance, Daw et al. [18] implemented a physics-guided loss in their ML model to predict lake temperatures across different depths. They aimed to ensure that the model predictions adhered to the monotonic relationship between water depth and density (i.e., denser water should reside at deeper depths). Subsequent works by Read et al. [73] and Jia et al. [30] refined this framework to incorporate energy conservation principles for temporal lake temperature predictions. Furthermore, in the field of LSM, there are also some studies that can be broadly classified under the PGML modeling strategy. Stanley et al. [81, 82] applied monotonic constraints to their ML models for LSM, in which a direction was assigned to model response for each input variable using prior knowledge based on their contribution to landslide risk (e.g., non-decreasing monotonicity between soil moisture and landslide risk). Wei et al. [97, 98] developed HPD models for LSM, wherein results from the infinite slope model were used as additional input features for ML models. This integration led to observed improvements in both model performance and generalization capability across different regions. Khabiri et al. [39] and Liu et al. [44] used physics-based models for negative sample selection. Their LSM models trained on datasets filtered through these physics-based models demonstrated higher performance and improved

interpretability. However, it should be noted that while these studies demonstrate advancements in model performance and interpretability, they remain either too restrictive or do not consistently guarantee physics consistency.

The present study proposed a PGML modeling framework that utilizes physics-based model regularizations to model landslide susceptibility. A well-documented debris flow inventory from a debris flow event in the Colorado Front Range was used to develop and evaluate the performance of the proposed PGML model. In the following sections, the study area and debris flow inventory are first described, followed by descriptions of the physics models, baseline ML models, PGML models, and the design of ML experiments. Last, the performance of these models was compared and evaluated. It should be noted that debris flow is a type of landslide [11]; in the subsequent discussions, the term debris flow is used to describe what actually occurred in the Colorado Front Range, whereas the term landslide is used in the context of LSM and downslope movement of soil in general. In addition, it should be noted that the landslide contributing factors used and the LSM models developed in this study are independent of dynamic, time-specific events such as rainfall. They aim to provide long-term, static assessments of areas inherently susceptible to landslides due to geological and environmental factors.

2 Study area and debris flow inventory

The generalizability of LSM models is fundamental for their effectiveness in estimating landslide susceptibility across diverse environmental conditions. To evaluate the generalization capability of LSM models, an area encompassing a wide range of ecological, topographical, and climatic conditions is necessary. The Colorado Front Range, a key segment of the Southern Rocky Mountains in North America, was selected as the study area due to its pronounced environmental diversity and historical prevalence of landslides.

Centrally located in the state of Colorado, this mountain range originated from the orogenic uplift caused by regional compression during the Laramide Orogeny in the Late Cretaceous to early Tertiary periods [17]. The range showcases a substantial elevation variation ranging between 1500 and 4300 m; it encompasses four major topographic elements and five distinct ecosystem zones [1, 14]. The vegetation density, soil development, and regolith production are dependent on the slope aspect, especially in the montane zones. The north-facing hillslopes are covered by dense coniferous forests and have more leached, colder soils compared to the south-facing slopes, which predominantly support grass with a few small shrubs [5, 47]. An ecoregion defines areas with similar types, quality, and quantity of environmental resources, such as biomes and topography [48]. Based on level IV ecoregions of the Conterminous U.S., the study area in the present study consists of seven distinct ecoregions, including Flat to Rolling Plains (FRP), Front Range Fans (FRF), Foothill Shrublands (FS), Crystalline Mid-Elevation Forests (CMEF), Crystalline Subalpine Forests (CSF), Sedimentary Mid-Elevation Forests (SMEF), and Alpine Zone (AZ) in the order of rising elevation. The location and extent of the study area and the ecoregion partitions within the study area are shown in Fig. 2.

The accuracy and extensiveness of landslide inventories are essential for the efficacy of data-driven LSM models. In this study, a comprehensive debris flow inventory with precise location was used to develop LSM models to discover connections between landslide contributing factors and landslide susceptibility within the study area. The inventory was mapped by Coe et al. [14] and is publicly available from the USGS Landslide Inventory, which can be accessed at https://www.usgs.gov/tools/us-landslide-inventory. It encompasses a broad range of debris flows mobilized from discrete sliding masses of colluvial soil (i.e., shallow landslides), spanning an area of 3430 km^2 in the northern portion of the Colorado Front Range. These debris flows are located across five ecoregions in the study area: FRF, FS, CMEF, CSF, and AZ. The origin of 97% of these shallow landslides can be traced back to open slopes (48%) or swales (49%), with channels only contributing to 3% of the initiations [14]. The inventory provides accurate coordinates for the initiation points of 1138 debris flows and 212 slides, which were mapped through field reconnaissance and the analysis of high-definition orthorectified satellite imagery [14]. This extensive debris flow inventory provides a rich dataset foundation to assess the performance of LSM models. The locations of these debris flows are presented in Fig. 2.

3 Landslide contributing factors

For statistically based LSM, landslide contributing factors need to be carefully selected to ensure they can reflect the effects of soil properties, hydrologic conditions, and terrain geometries that correspond to landslide formation. Given the data accessibility and the nature of landslides that occurred in the study area, nine landslide contributing factors were selected for the present study, ensuring a relevant and robust data foundation for training ML models to predict landslide susceptibility accurately [68]: elevation, slope, aspect, topographic wetness index (TWI), normalized difference vegetation index (NDVI), sand content, clay content, bulk density, and field capacity.

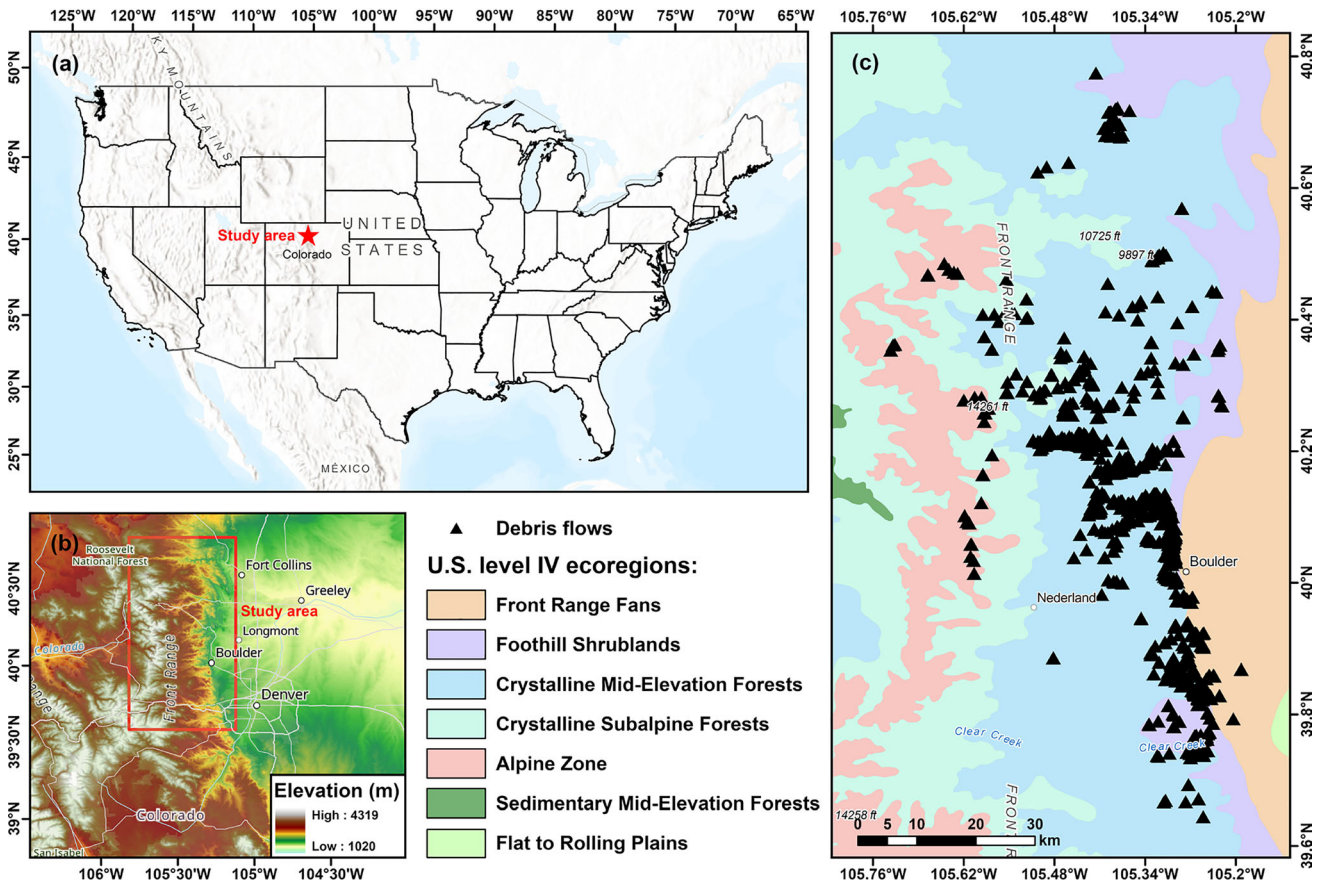

Fig. 2 Summary of the study area and debris flow locations: **a** continental United States; **b** State of Colorado; **c** overview of the study area, including debris flow locations and ecoregion partitions

Among these factors, elevation is based on the high-resolution digital elevation model (DEM) provided by the 3D Elevation Program (3DEP) [86] with a spatial resolution of 10 m, which is a reasonable value to represent the size of a typical landslide detachment area while also capturing local topographic gradients. Slope and aspect were calculated from the DEM, representing the local gradient of the hillslope and its cardinal directions. TWI was derived from the DEM to quantify topographic control on hydrological processes, which is calculated as:

$$\mathrm{TWI} = \ln\left(\frac{A_s}{\tan\beta}\right) \tag{1}$$

where A_s is the specific catchment area and β is the hillslope angle. NDVI was used to represent surface vegetation coverage and was calculated based on near-infrared and red bands from Landsat-7 satellite images:

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{R}}{\mathrm{NIR} + \mathrm{R}} \tag{2}$$

where NIR is the near-infrared portion and R is the red portion of the electromagnetic spectrum. In the present study, Landsat-7 satellite images taken within half a year before the September storm event were used to calculate the average NDVI for the study area. The soil information (i.e., sand/clay content, soil bulk density, and field capacity) was obtained from SoilGrids [34], which provides a global estimation of a wide range of soil, land cover, hydrology, geology, climate, and relief characteristics with a spatial resolution of 250 m. These soil properties are provided for six standard depths up to 200 cm produced by ML algorithms trained on global soil profiles. The present study used the depth-weighted average to obtain soil properties at each location. These landslide contributing factors were stored as georeferenced raster images. The respective values at various locations can be retrieved by extracting the corresponding pixel values from these images. Figure 3 presents a visualization of these landslide contributing factors for the study area. All data used in this study, including landslide inventory and landslide contributing factors, are summarized in Table 1. To facilitate subsequent modeling and analysis, all the landslide contributing factors were resampled to the same resolution as the DEM (i.e., elevation).

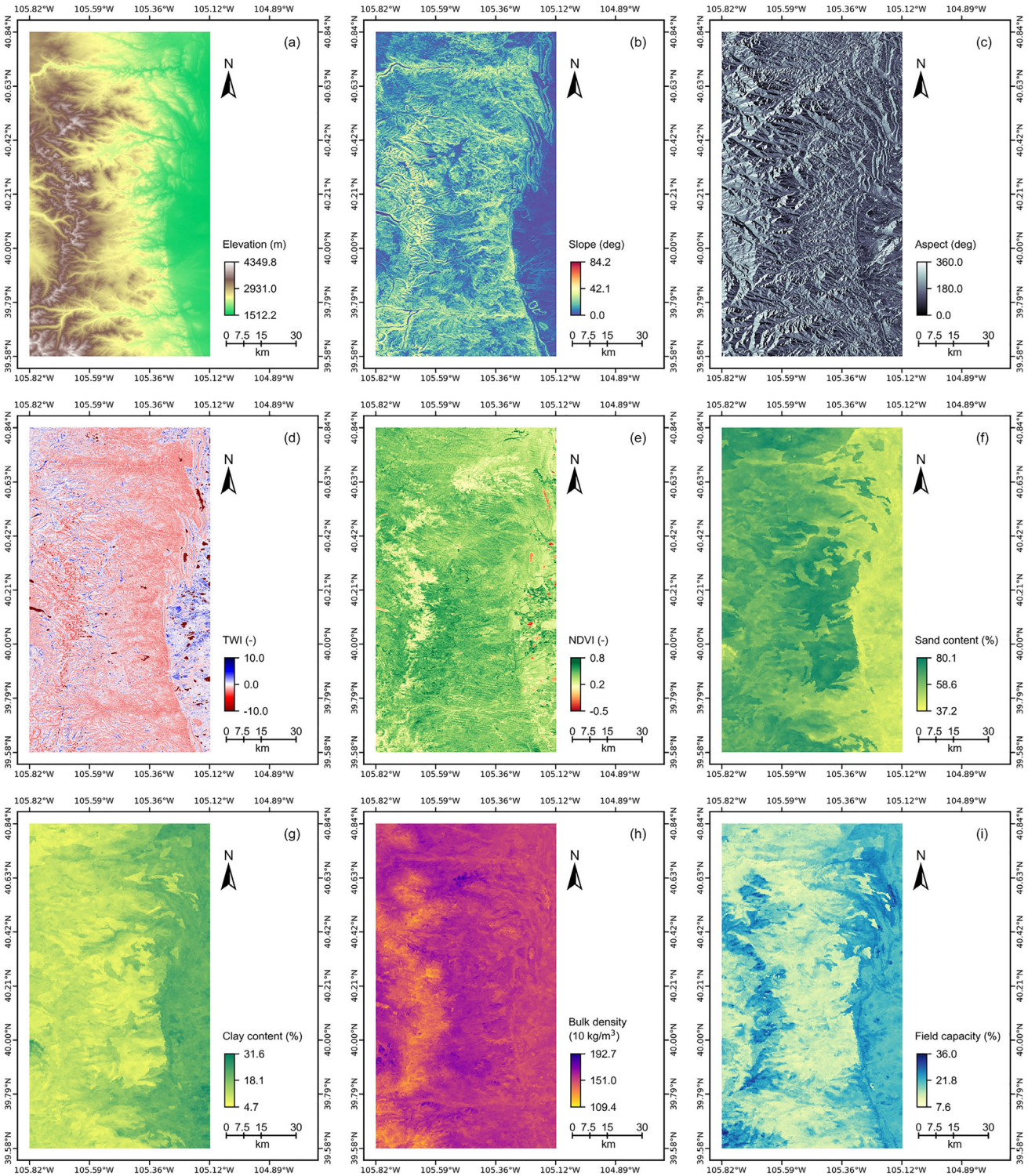

Fig. 3 Thematic maps of landslide contributing factors for the study area: **a** elevation; **b** slope; **c** aspect; **d** TWI; **e** NDVI; **f** sand content; **g** clay content; **h** bulk density; **i** field capacity

Table 1 Summary of data used in the present study

Data	Unit	Type	Resolution	Data source
Landslide inventory	–	Point	–	USGS landslide inventory
Elevation	m	Raster	10 m	USGS 3DEP
Slope	deg	Raster	10 m	Derived from elevation
Aspect	deg	Raster	10 m	Derived from elevation
Topographic Wetness Index (TWI)	–	Raster	10 m	Derived from elevation
Normalized difference vegetation index (NDVI)	–	Raster	30 m	Landsat 7
Sand content	%	Raster	250 m	SoilGrids
Clay content	%	Raster	250 m	SoilGrids
Bulk density	10 kg/m^3	Raster	250 m	SoilGrids
Field capacity	%	Raster	250 m	SoilGrids

4 Physics-based model

In this study, the infinite slope model was selected as the physics-based model owing to its suitability for predicting shallow landslides and its simplicity and compatibility with Geographic Information Systems (GIS) for grid-based analysis on regional scales [101]. The factor of safety (FoS) for the infinite slope model is calculated as the ratio of the soil shear strength to the shear stress imposed by the slope material:

$$\mathrm{FoS} = \frac{c + (\gamma d - \gamma_w h)\cos^2\beta\tan\phi}{\gamma d \sin\beta\cos\beta} \tag{3}$$

where h represents the height of the groundwater table, d denotes the depth of the potential sliding plane, β is the slope angle. The terms γ and γ_w correspond to the unit weight of soil and the unit weight of water, respectively. And ϕ and c are the soil friction angle and cohesion, respectively. For simplicity and to ensure hydrological consistency, submerged slopes with seepage parallel to the slope were assumed when calculating the FoS (i.e., $h = d$). This approach represents a conservative scenario and allows for a uniform assessment of landslide susceptibility across the study area rather than modeling the response to specific rainfall events. The input parameters for the infinite slope model were estimated based on landslide contributing factors using empirical relationships and previous studies in the study area [65]. For example, β is based on the hillslope angle from DEM. d is estimated based on an elevation-dependent relationship [77] as:

$$d_i = d_{\max} - \frac{z_i - z_{\min}}{z_{\max} - z_{\min}}(d_{\max} - d_{\min}) \tag{4}$$

where d_i and z_i are local soil layer thickness and terrain elevation, respectively; $z_{\min}$ and $z_{\max}$ are the minimum and the maximum terrain elevation, respectively; and $d_{\min}$ and $d_{\max}$ are the minimum and the maximum soil layer thickness, respectively. The elevation-dependent soil thickness relationship assumes that the elevation and soil thickness are inversely related. This assumption suggests that soil erosion increases with elevation, resulting in shallower soils, while sedimentation occurs at lower elevations, forming thick colluvial and alluvial soils [40]. As summarized in several previous studies (e.g., [2, 9, 50]), the soil thickness in the study area is generally thin and exposed bedrock is observed in some areas. The estimated soil thickness for landslide source areas clustered between 0.4 and 1.3 m, with a few thicker than 2.0 m on less steep slopes, based on the reconnaissance by Coe et al. [14]. In the present study, $d_{\min} = 0.1$ m and $d_{\max} = 1.5$ m were used to reflect the decreasing trend of soil layer thickness as elevation increases for the study area. Tiwari and Marui [84] reported soil residual friction angle versus clay contents for 82 natural disasters associated with slope stability problems. This study used curve fitting ($R^2 = 0.49$) to correlate soil clay content with residual friction angle based on these 82 samples:

$$\tan\phi = -0.0978\ln(\text{clay content}) + 0.575 \tag{5}$$

For estimating c, the soil was assumed to be cohesionless; however, an apparent cohesion due to root reinforcement in slope stability was considered. The values of c were estimated by applying a linear transformation to the full spectrum values of NDVI using the following equation [31]:

$$c = c_{\min} + c_{\text{int}} \times \frac{\mathrm{NDVI} + 1}{2} \tag{6}$$

where $c_{\min}$ and c_{int} are constants controlling the minimum value and range of the cohesion. In the present study, $c_{\min} = 0$ kPa and $c_{\text{int}} = 8$ kPa produce apparent cohesions within the same range as previous studies (e.g., [9, 50]). The soil unit weight is based on bulk density and field capacity from SoilGrids. Assuming the initial volumetric

water content of the soil is at the field capacity, which is the moisture content above which the soil layer is drained by gravity. Then, the saturated unit weight of soils can be calculated as:

$$\gamma_{\text{sat}} = \gamma_{\text{b}} + (n - \vartheta_{\text{FC}})\gamma_w \tag{7}$$

where γ_{b} is the bulk density; ϑ_{FC} is the field capacity, which is reported as the volumetric water content [34]; and n is the soil porosity, which can be estimated using the following equation:

$$n = 1 - \frac{\gamma_{\text{b}}}{\gamma_w G_{\text{s}}} \tag{8}$$

where G_{s} is the specific gravity of soil particles and is assumed to be 2.65 in this study.

5 ML model

This study used the multi-layer perceptron neural network (MLPNN) to develop models for predicting landslide susceptibility for the study area. MLPNNs consist of multiple layers of interconnected nodes or neurons and are capable of learning nonlinear relationships between input features and target outputs, making them suitable for modeling complex natural processes. In this study, the MLPNN was designed to have three hidden layers, each containing 24 neurons; the rectified linear unit (ReLU) activation function [3] was applied in the hidden layers, providing better training efficiency and mitigating the vanishing gradient problem. The sigmoid activation function was applied to the output layer for the binary classification task, and the binary cross-entropy loss function was employed to measure the error between the network's predictions and ground truth labels. The Adam optimization algorithm was chosen for optimizing the weights and biases for the network with a learning rate of 0.001. In addition, all the models were trained for 50 epochs with a batch size of 16 during the model development procedure. A detailed description of MLPNN can be found in Kuhn and Johnson [37] and Goodfellow et al. [27].

6 PGML framework

Traditional machine learning (ML) models (e.g., MLPNN) often struggle to accurately represent complex scientific relationships derived directly from data, particularly when the available training data are insufficient [94]. Recently, researchers have adopted physics-guided loss functions to address challenges encountered by traditional ML models for various applications (e.g., [30, 66, 73]). The PGML framework integrates domain-specific scientific relationships into the loss functions during the model training process, serving as regularization terms that guide the model toward physical consistency and enhanced performance. A generic form of the physics-guided loss function L_{pg} can be expressed as:

$$L_{\text{pg}} = \underbrace{L_{\text{data}} + \lambda_{\text{r}} R}_{\text{standard loss for ML models}} + \underbrace{\lambda_{\text{phy}} L_{\text{phy}}}_{\text{physics-based loss}} \tag{9}$$

where L_{data} represents the supervised error between the prediction and the ground truth, R is the regularization loss that optimizes model simplicity, and L_{phy} is the physics-based loss that measures the consistency of model predictions with respect to domain-specific scientific relationships. These three terms optimize three aspects of model performance in terms of accuracy, simplicity, and consistency. Parameters λ_{r} and λ_{phy} are hyperparameters that control the respective weights of R and L_{phy} in the physics-guided loss function, respectively. Note that the first two terms are the standard loss for traditional ML models. For slope stability analysis, there is a generally accepted monotonic relationship between the landslide susceptibility and the FoS. The FoS measures the ratio between the resisting force and the driving force along a potential failure surface. A higher FoS value generally indicates decreased landslide susceptibility, establishing a foundational physics-based relationship that the model can learn and respect. However, ML models trained solely with data may not reflect this physical relationship, yielding less scientifically interpretable predictions. Therefore, guiding the model toward physics consistency is desirable during the training process. The monotonic relationship between landslide susceptibility based on model predictions $\hat{y}$ and FoS values for any two samples can be expressed as:

$$\hat{y}_1 - \hat{y}_2 \leq 0 \text{ if } \text{FoS}_1 \geq \text{FoS}_2 \tag{10}$$

If samples in each training batch are sorted based on their FoS values in a descending manner (i.e.,$FoS_i \geq FoS_{i+1}$), a difference in model predictions can be computed for any pair of sequential samples as:

$$\Delta\hat{y} = \hat{y}_{i+1} - \hat{y}_i \tag{11}$$

Hence, a negative value of $\Delta\hat{y}$ can be considered a violation of physics. Similar to the approach proposed by Daw et al. [18] and Pei et al. [66], a physics-based loss term L_{phy} that measures the average value of these violations of physics (i.e., physics inconsistency) can be expressed as:

$$L_{\text{phy}} = \frac{1}{n-1}\sum_{i=1}^{n-1} \max(-\Delta\hat{y}, 0) \tag{12}$$

where n is the number of training samples. Landslide susceptibility prediction can be considered a binary

classification problem. The binary cross-entropy loss [10] is typically used as the standard loss and can be expressed as:

$$L_{\text{data}} = \frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i}) \tag{13}$$

where y and $\hat{y}$ are labels and model predictions, respectively. It should be noted that various regularization methods have been extensively explored in the data science community to implement different measures of regularization; however, these approaches often neglect the physical plausibility of trained models, potentially leading to predictions that lack physical consistency [38]. Therefore, the model complexity loss may inadvertently dilute the emphasis on physical consistency and was not considered in this study. Based on L_{data} and L_{phy}, a complete form of the physics-guided loss function for this study can be expressed as:

$$L_{\text{pg}} = L_{\text{data}} + \lambda_{\text{phy}} L_{\text{phy}} \tag{14}$$

For the PGML model, the physics-based loss term in Eq. (14) steers the model toward a monotonic relationship between the model prediction and the calculated FoS based on the physics-based model. A detailed illustration of the PGML model training process can be found in Algorithm 1. The PGML model structure and hyperparameters remain the same as the baseline MLPNN model except for the loss function, and the same validation procedure was used to evaluate the model performance, which will be discussed in the following sections.

Algorithm 1 PGML model training process

Data: Training data, initial model parameters, hyperparameters λ_{phy}
Result: Trained model with integrated physics-guided principles
for *each epoch* **do**
  for *each batch* **do**
    Initialize batch loss to 0
    $\hat{y} \leftarrow \text{model.forward}(X_{\text{batch}})$
    $L_{\text{data}} \leftarrow -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$
    $\text{indices} \leftarrow \text{argsort}(\text{FoS}_{\text{batch}}, \text{descending=True})$
    $\hat{y}_{\text{sorted}} \leftarrow \hat{y}[\text{indices}]$
    $\text{FoS}_{\text{sorted}} \leftarrow \text{FoS}_{\text{batch}}[\text{indices}]$
    $L_{\text{phy}} \leftarrow 0$
    for $i \leftarrow 1$ **to** $n - 1$ **do**
      if $FoS_{sorted}[i] \geq FoS_{sorted}[i + 1]$ **then**
        $\Delta\hat{y} \leftarrow \hat{y}_{\text{sorted}}[i] - \hat{y}_{\text{sorted}}[i + 1]$
        if $\Delta\hat{y} > 0$ **then**
          $L_{\text{phy}} \leftarrow L_{\text{phy}} + \Delta\hat{y}$
        end
      end
    end
    $L_{\text{phy}} \leftarrow \frac{1}{n-1} L_{\text{phy}}$
    $L_{\text{pg}} \leftarrow L_{\text{data}} + \lambda_{\text{phy}} \times L_{\text{phy}}$
    $\text{model.backward}(L_{\text{pg}})$
  end
end

7 Model performance evaluation

In this study, the receiver operating characteristic (ROC) curve was used to evaluate the model performance. The ROC curve is a two-dimensional graphical representation that illustrates the performance of a classification model by depicting the relationship between the false positive rate (FPR) and the true positive rate (TPR) at various classification thresholds. The area under the ROC curve (AUC) can be calculated, offering an aggregate measure of model performance. The AUC score is a prominent single-value metric utilized in classification model evaluations, quantifying the model's proficiency in distinguishing between two classes. A model that merely predicts at random would yield an AUC of 0.5, whereas a model with perfect classification capabilities would achieve an AUC of 1.0. Additionally, the confusion matrix reports four possible outcomes of model predictions at a given classification threshold: (1) true positive (TP), which represents correctly predicted landslide samples; (2) true negative (TN), which represents the number of correctly predicted non-landslide samples; (3) false positive (FP), which denotes misclassified landslide samples; and (4) false negative (FN), which denotes misclassified non-landslide samples. After identifying the optimal classification threshold from the ROC curve, four commonly used model performance evaluation metrics can be calculated based on the confusion matrix: Accuracy, Precision, Recall, and F_1. These performance evaluation metrics, including FPR and TPR, can be calculated using the following equations:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{15}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{16}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{17}$$

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{18}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{19}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{20}$$

The Accuracy metric provides a holistic view of the model's predictions, representing the fraction of all correct predictions. Precision focuses on the model's ability to correctly identify landslides, while Recall emphasizes the model's capability to capture all actual landslide events. Note that the F_1 score provides an aggregate measure of the model performance score by calculating the harmonic mean of Precision and Recall. Besides the above-

mentioned metrics, a distinctive feature of the evaluation framework used in this study is the incorporation of a metric assessing the model's adherence to physics principles, termed Physics Inconsistency (PI). As described in the previous section, the calculation of PI is the same as L_{phy} using Eq. (12). In this study, the calculated FoS values based on the physics-based model were used as the reference (i.e., ground truth) to evaluate PI and guide model training. In light of the evaluation metrics outlined above, this study employed six key metrics for a comprehensive model performance evaluation, including Accuracy, Precision, Recall, F_1, AUC, and PI.

8 Dataset preparation

In the context of LSM, employing ML techniques formulates the problem into a binary classification task. In this study, the 1350 mapped debris flow locations in the study area and their corresponding values of landslide contributing factors were used as positive samples. Based on information reported in the USGS landslide inventory, these debris flow locations were mapped from the aftermath of an intense storm event in September 2013 [14]. Figure 4 presents a map of cumulative rainfall during the storm event for the study area. Figure 4 shows that the mapped debris flow locations were all in the areas with an cumulative rainfall more than 120 mm; a plausible interpretation of this phenomenon is that it takes more than 120 mm cumulative rainfall to fully saturate the slopes in the study area to initiate debris flows. Negative samples were strategically selected from regions experiencing cumulative rainfall exceeding 120 mm to force the ML models to learn why debris flows were not triggered despite having sufficient rainfall. These negative samples were also positioned at least 100 m away from any known debris flow locations to ensure they represent true non-debris flow areas. Figure 4 shows the distribution of debris flows and non-debris flows within the study area, overlaid on a map of cumulative rainfall for the September 2013 storm event. In Fig. 4, red triangles represent debris flow locations, black dots indicate non-debris flow areas, and isolines represent different levels of cumulative event rainfall. Besides sample selection, ensuring a balanced dataset is crucial for preventing model bias toward the predominant class and enhancing predictive accuracy; an equivalent number of negative and positive samples were drawn for each ecoregion, resulting in a dataset of 2700 samples for developing LSM models. Each sample was characterized by nine input features and a singular output/target (i.e., 1 for debris flow or 0 for non-debris flow). Table 2 presents the summary of the dataset size for each ecoregion and shows that most debris flows occurred in the CMEF, whereas very few occurred in the CSF and AZ. Given the substantial variations in the values of landslide contributing factors, as shown in Fig. 3, standardization of each feature was used in this study to facilitate effective model training.

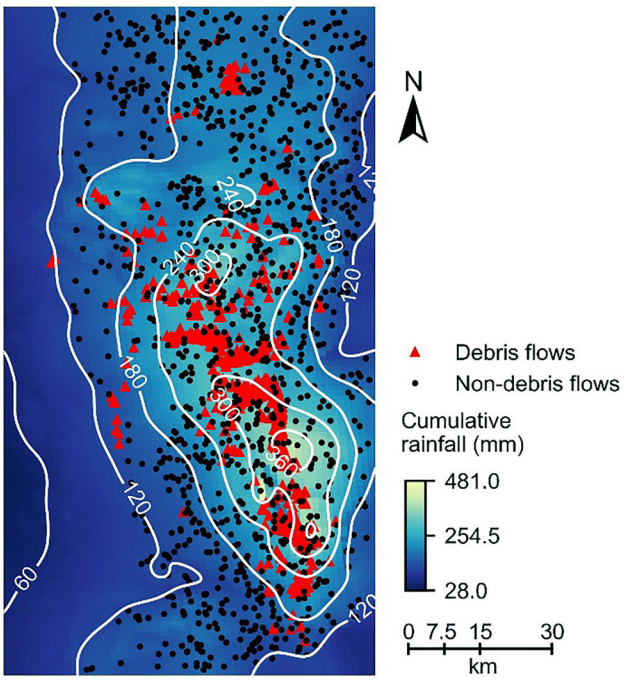

Fig. 4 Spatial distribution of debris flows and non-debris flows within the study area overlaid on a map of cumulative rainfall for the September 2013 storm event

Additionally, the t-distributed stochastic neighbor embedding (t-SNE) was used to examine the distribution of input datasets in this study. t-SNE is a dimensionality reduction technique that creates a low-dimensional representation of high-dimensional data [87]. A t-SNE plot visually displays the structure and relationships within the data by grouping similar data points close together and dissimilar data points further apart in a two- or three-dimensional space. This visualization can help reveal patterns, clusters, and potential outliers, making it a valuable

Table 2 Summary of dataset size for each ecoregion

Ecoregion	Number of positive, negative, and total samples
Front Range Fans (FRF)	127, 127, 254
Foothill Shrublands (FS)	192, 192, 384
Crystalline Mid-Elevation Forests (CMEF)	976, 976, 1952
Crystalline Subalpine Forests (CSF)	27, 27, 54
Alpine Zone (AZ)	28, 28, 56

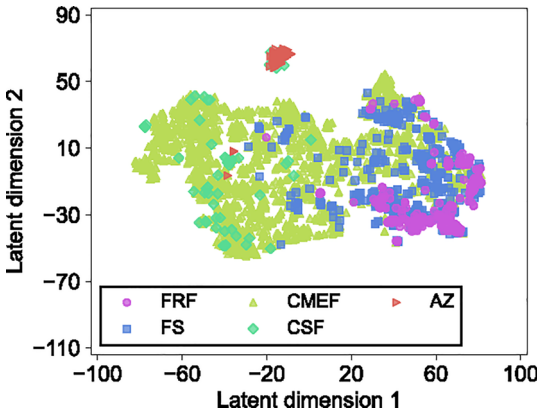

Fig. 5 Two-dimensional t-SNE visualization of input datasets

tool for exploratory data analysis and assessing the quality of features or representations learned by ML models. Figure 5 presents the two-dimensional t-SNE visualization of input datasets based on nine landslide contributing factors for all five ecoregions. As shown in Fig. 5, the input datasets exhibit different distributions among ecoregions, especially for the AZ, where data distribution significantly differs from the other ecoregions. The difference in data distribution can be attributed to the fact that these ecoregions contain unique ecological and geological characteristics distinct from each other. For example, the AZ is characterized by steep slopes and shallow soil to exposed rock, which is significantly different than the FRF.

9 Model development

The model development procedure in this study includes data preparation, model training and performance evaluation, ensemble model formulation, and final deployment for generating landslide susceptibility maps for the study area. The Python package PyTorch [62] was used to develop MLPNN and PGML models. To facilitate the illustration of the model development procedure, Fig. 6 shows a workflow chart for developing the PGML.

During the development of LSM models, ensuring that models possess strong generalization capabilities across diverse geographical regions is crucial. Given the inherent differences between hillslopes and ecoregions, it is essential to assess how well these models can extrapolate to areas not represented in the training data, which requires a robust validation technique that effectively evaluates the model's generalization performance.

Cross-validation (CV) is a commonly used model validation technique to evaluate ML model performance, and it is particularly beneficial when working with limited data. In this process, the dataset is divided into k subsets (i.e., folds). Each subset serves as a validation set, while the model is trained on the remaining $k-1$ subsets (or folds). This process repeats k times, each with a different subset serving as the validation fold. The final model performance is derived by averaging the performance metrics across all k iterations, ensuring a robust assessment of the model performance and stability on unseen data.

However, generic CV methodologies based on random sampling, such as random CV, often assume that data samples are independently and identically distributed. This assumption can result in overly optimistic performance estimates when applied to data with inherent temporal, spatial, hierarchical, or phylogenetic structures commonly found in fields like ecology and geotech/geoscience. To address this, it is preferable to sample data into blocks that reflect their intrinsic structure (e.g., spatial autocorrelation in landslide data [6, 8, 69, 80]). This approach helps ensure the training and validation datasets are independent and more accurately represent the complexities of the data. [71]. As an alternative, spatial CV is a method used to evaluate the performance of predictive models in geospatial applications (e.g., [51, 71]), including landslide detection and prediction (e.g., [41, 69, 79]), which gives a more realistic assessment of the model performance by ensuring that the training and validation sets are spatially independent. In spatial CV, the dataset is divided into multiple spatially disjoint subsets (folds); for each fold, the model is trained on the remaining folds and then tested on the target fold. The extent of ecoregions was used in this study as the dataset partition strategy for spatial CV, which aims to evaluate the model's generalization capability across heterogeneous environments. It should be noted that the spatial CV framework used in this study is a generic approach applicable for evaluating various models, including both physics-based and data-driven (e.g., ML) models. Moreover, the entire spatial CV procedure was repeated five times with varying random seeds to accommodate and evaluate data and model uncertainties.

After model training and validation, one spatial CV was randomly chosen from the five spatial CV repetitions. The five candidate models produced from this single spatial CV were then aggregated using the average ensembling method [32] to generate a landslide susceptibility map for the entire study area. This method not only enhances the robustness and generalization of the final model by utilizing models trained on different ecoregion blocks but also retains the structured nature of different ecoregions in the individual models. The preservation of ecoregional structures in each individual model provides a more accurate estimation of the prediction uncertainty for the ensemble model across the entire study area. This approach can effectively reflect model uncertainty due to geographical and environmental diversity, providing a way to evaluate

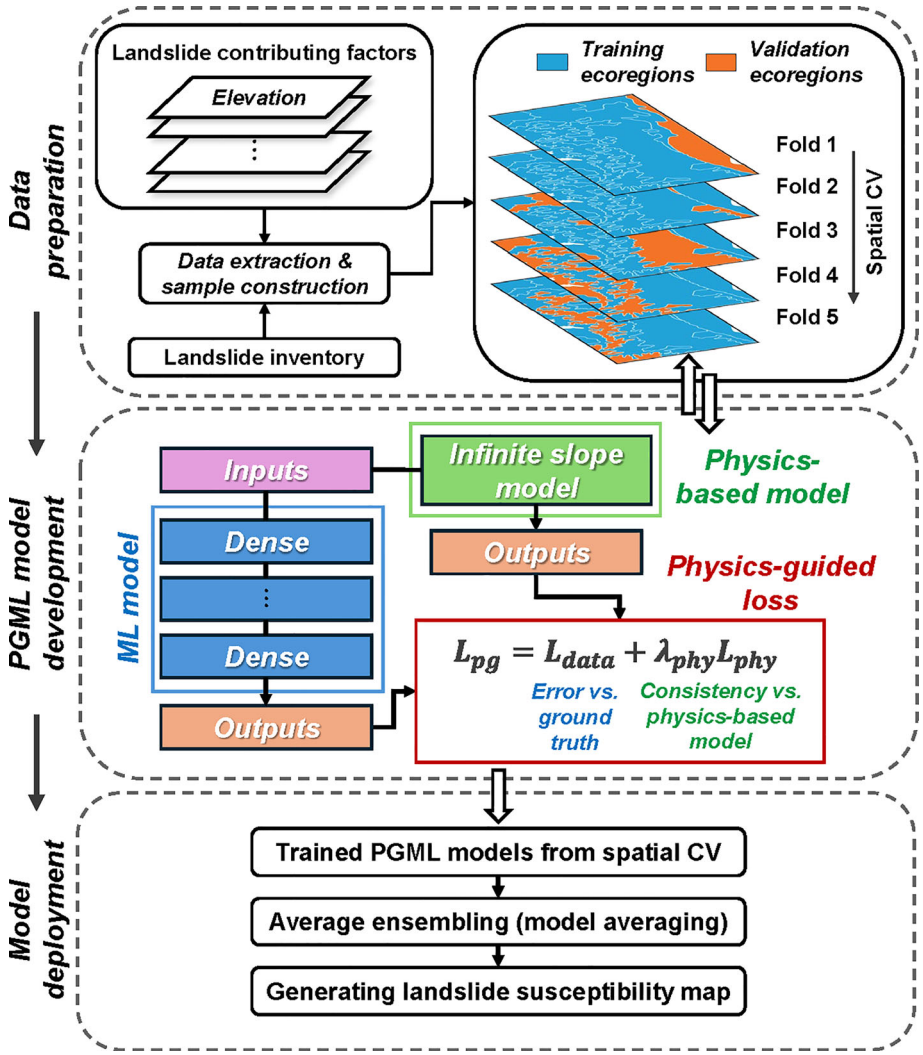

Fig. 6 Schematic overview of the PGML model development process for LSM

the reliability of the LSM model predictions in varied landscapes.

10 Performance of physics-based model

Figure 7 presents the landslide susceptibility map based on the physics-based model overlaid with mapped debris flow locations, where the calculated FoS at each location is shown as a contour for the study area. As shown in Fig. 7, the concurrence of predicted low FoS areas and mapped debris flow locations suggests that the geotechnical domain knowledge, such as the infinite slope model and estimated soil parameters, can effectively estimate regional-scale debris flow risks when appropriately applied. The performance of the physics-based model was evaluated using the same spatial CV procedure as ML models based on the dataset summarized in Table 2 to assess its performance on each ecoregion and its generalization capability across the ecoregions. It is important to note that, unlike ML models, the physics-based model does not require training, and its performance is independent of dataset variability. Figure 8 presents the ROC curves for the physics-based model for each ecoregion. These curves are generated by converting the continuous output of the physics-based model (i.e., FoS) into binary classifications through the application of various threshold values. By mapping through a range of FoS threshold values, continuous measures are transformed into discrete categories that indicate susceptibility. Samples with a FoS below the threshold are classified as susceptible to landslides (i.e., 1), while those above are classified as not susceptible (i.e., 0). The TPR and FPR are then calculated to construct the ROC curve. It is noteworthy that this method of generating ROC curves by mapping through thresholds is the same as the approach used in ML. The primary difference is the nature of the

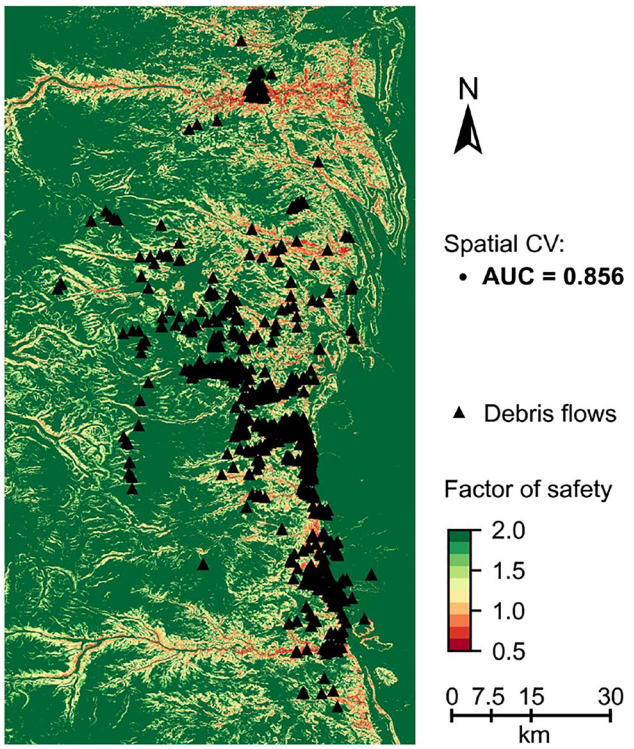

Fig. 7 Landslide susceptibility map for the study area based on the physics-based model

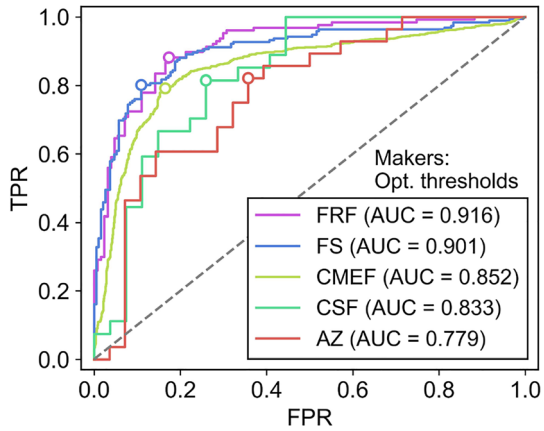

Fig. 8 ROC curve for the physics-based model based on spatial CV (diagonal line represents a random classifier)

thresholds used: physics-based models apply thresholds based on FoS values, whereas ML models typically employ probability thresholds ranging from 0 to 1. As shown in Fig. 8, the physics-based model generally performs well in distinguishing between landslide and non-landslide classes in these ecoregions, with an average AUC score of 0.856. However, the performance of the physics-based model varies across different ecoregions. For example, the physics-based model achieved excellent performance for the FRF and FS ecoregions with AUC scores above 0.9, whereas the performance for the AZ was relatively low with an AUC score below 0.8. These discrepancies may be attributed to the empirical relationships used to determine input parameters (e.g., Eqs. (4) to (6)) and the applicability of the infinite slope model for different ecoregions. For example, the infinite slope model may not be applicable to areas with exposed rock in the AZ, resulting in a low AUC score for the ecoregion.

Based on reviews of case histories, Bowles [4] summarized that a slope is generally deemed unstable and prone to failure if FoS < 1.07, a moderate risk of failure if $1.07 <$ FoS < 1.25, and relatively stable if FoS > 1.25. Thus, FoS = 1.25 was often used as the default threshold for the binary classification of slope stability for physics-based models. The optimal classification threshold can be obtained from the ROC curve as the FoS value corresponding to the point on the ROC curve closest to the top-left corner (0, 1) that represents perfect classification. These points are marked on Fig. 8 as circles and the corresponding optimal threshold FoS values are reported in Table 3. Figure 9 compares the classification performance for the physics-based model using optimal thresholds determined from the ROC curves for each ecoregion versus a default threshold of FoS = 1.25. As shown in Fig. 9, the optimal thresholds generally yield better classification performance with less variation compared to the fixed threshold value of 1.25, which can be attributed to the fact that the optimal threshold considers the trade-off between different types of misclassifications and is adjusted according to the specific characteristics of the data and task. Consequently, a more balanced and improved performance can be achieved using the optimal threshold FoS value.

Table 3 presents detailed classification scores based on the optimal threshold value for each ecoregion. In addition, Table 3 also presents classification scores based on the five-fold random CV, for which stratified random sampling was used for dataset partition, and non-debris flow samples were randomly selected within the entire study area, disregarding the ecoregions. As shown in Table 3, the optimal classification threshold varies among these ecoregions, which indicates that the generalized empirical relationships used to determine input parameters and the infinite slope model may not produce optimal values for all ecoregions, and region-specific analysis should be used for calibrating the physics-based model. Notably, Table 3 also shows that the random CV yielded better performance scores than the spatial CV. However, random CV ignores spatial dependencies in the dataset and may not provide a rigorous and realistic assessment of a model's ability to generalize to unseen data or from one ecoregion to another, which will be further discussed.

Table 3 Validation performance of the physics-based model across ecoregions

Ecoregion	Accuracy	Precision	Recall	F_1	AUC	Optimal classification threshold
FRF	0.850	0.868	0.827	0.847	0.916	1.821
FS	0.841	0.810	0.891	0.849	0.901	1.259
CMEF	0.813	0.800	0.835	0.817	0.852	1.110
CSF	0.759	0.769	0.741	0.755	0.833	1.482
AZ	0.714	0.750	0.643	0.692	0.779	1.815
Avg	0.796	0.799	0.787	0.792	0.856	1.498
Random CV	0.833	0.837	0.828	0.832	0.885	1.196

11 Performance of MLPNN model

In this study, the spatial CV procedure was repeated five times with different random seeds. Figure 10 presents ROC curves for the MLPNN model for each ecoregion based on spatial CV, with the lines representing mean values and shaded areas representing standard deviations. As shown in Fig. 10, the performance of MLPNN models is generally worse than the physics-based model (see Fig. 8), with an average AUC score of 0.805. It should be noted that spatial CV tests the model's generalization capability on distinct environments from one region to another, using the extent of ecoregions as the sampling strategy. The relatively low AUC score of the MLPNN model during validation can be attributed to the fact that the traditional ML models (i.e., the MLPNN model) rely significantly on patterns in the training data, limiting their adaptability to new scenarios. In addition, unlike physics-based models grounded by well-established physical rules, traditional ML models do not have inherent rules or principles guiding their predictions. They rely solely on identifying patterns in data, and the model's performance can degrade if these patterns do not hold in new data; in other words, applying the ML model trained based on the data from one ecoregion to another may have significantly worse performance. In particular, this is observed in the case of the AZ ecoregion, where the MLPNN model had a low validation AUC score of 0.580, similar to the performance of a random classifier, which can be expected as the AZ ecoregion is drastically different from the rest of the ecoregions (see Fig. 5). In addition, Fig. 10 also shows significant variation between spatial CV repetitions, which can be attributed to uncertainties in data sampling and model development. Figure 11 compares classification performance between the

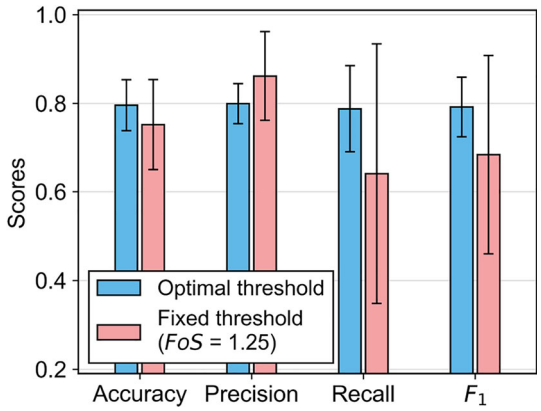

Fig. 9 Effect of classification thresholds on the physics-based model performance based on spatial CV

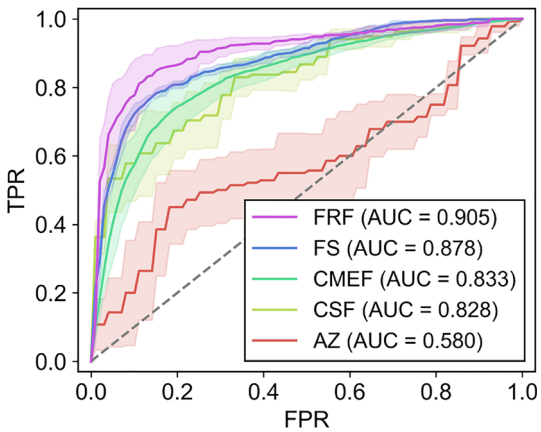

Fig. 10 ROC curve for the MLPNN model based on spatial CV

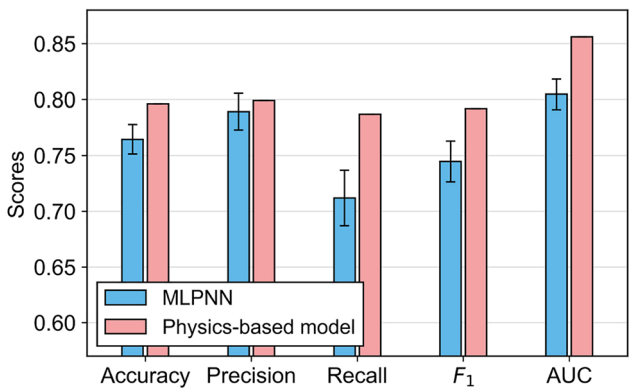

Fig. 11 Comparison of classification performance between the MLPNN and the physics-based model based on spatial CV

Table 4 Validation performance of the MLPNN model across ecoregions

Ecoregion	Accuracy	Precision	Recall	F_1	AUC	PI	Optimal classification threshold
FRF	0.854	0.861	0.844	0.852	0.905	0.133	0.117
FS	0.818	0.835	0.793	0.813	0.878	0.222	0.454
CMEF	0.777	0.786	0.762	0.774	0.833	0.227	0.747
CSF	0.752	0.791	0.696	0.737	0.828	0.171	0.196
AZ	0.621	0.673	0.464	0.547	0.580	0.139	0.169
Avg	0.765	0.789	0.712	0.745	0.805	0.178	0.337
Random CV	0.867	0.873	0.859	0.866	0.932	0.194	0.486

MLPNN and the physics-based model based on their corresponding optimal classification thresholds. It is evident from Fig. 11 that the MLPNN model generally underperforms when compared to the physics-based model in terms of binary classification performance. Detailed classification scores and optimal classification thresholds for the MLPNN model across validation ecoregions are shown in Table 4. It can be noted from Table 4 that the optimal thresholds for MLPNN models across validation ecoregions exhibit significant variations, which are markedly different from the standard 0.5. This variation is due to the distinct characteristics and data distributions in each ecoregion, which influence the model's TPR and FPR differently. Therefore, it is necessary to conduct region-specific threshold adjustments to accurately reflect the unique environmental and geological factors influencing landslide susceptibility in different areas.

Table 4 also presents the corresponding results using the fivefold random CV and PI scores. As shown in Table 4, the MLPNN model based on random CV showed significantly higher classification scores than those based on spatial CV. However, these numbers can be misleading as spatial autocorrelation was ignored during model training and evaluation, and the actual generalization capability of the model might be low, which is evident from the spatial CV results. By comparing Tables 3 and 4, it is also evident that the MLPNN showed better performance scores than the physics-based model using random CV. This indicates that the MLPNN model can effectively extract features from training data and perform well on validation data with similar distributions; however, its generalization performance could be less reliable than the physics-based model. Additionally, the MLPNN model showed a PI score of 0.178 for the spatial CV and 0.194 for the random CV, respectively. These PI scores indicate that pure data-driven MLPNN models may produce results that disobey the underlying physical relationship contributing to landslide susceptibility, such as monotonic relationships between landslide susceptibility and FoS, which will be further discussed in the following sections.

After MLPNN models were trained and evaluated, one spatial CV repetition was randomly chosen, and the five candidate models produced from this repetition were aggregated to create an ensemble model, referred to as the MLPNN ensemble model, using the average ensembling approach described in the Model Development section. This MLPNN ensemble model was then used to generate the landslide susceptibility map depicted in Fig. 12, which is essentially a contour of ML model prediction/output, ranging from 0 (non-debris flow locations) to 1 (debris flow locations); hence, the model output can be interpreted as landslide susceptibility. As shown in Fig. 12, the predicted areas of high landslide susceptibility closely align with the mapped locations on the eastern side of the

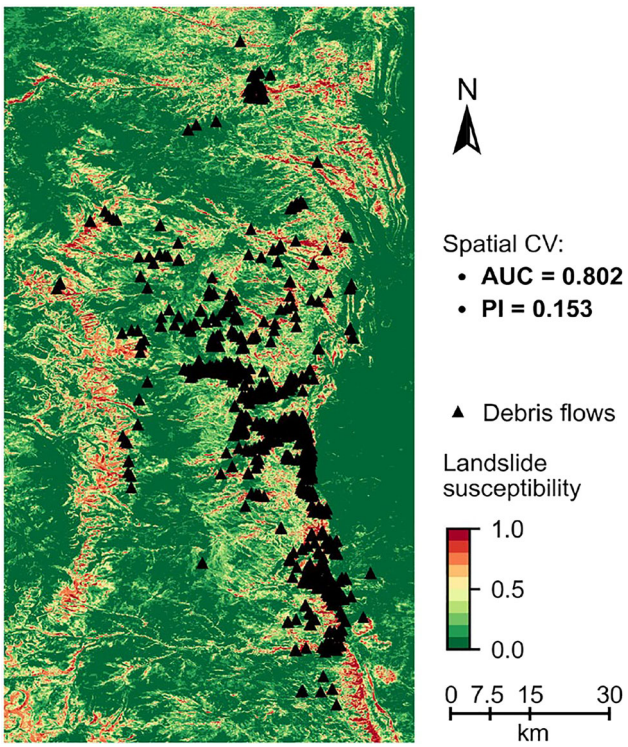

Fig. 12 Landslide susceptibility map for the study area based on the MLPNN ensemble model from one spatial CV repetition

study area (i.e., FRF, FS, and CMEF ecoregions) where mapped debris flows are clustered. However, despite limited numbers of mapped debris flow locations, the MLPNN ensemble model predicted high landslide susceptibility for the western part of the study area (i.e., CSF and AZ ecoregions) with steep slopes and shallow soils. This overprediction can be attributed to the model's reliance on dominant features such as slope, which were heavily weighted due to their strong correlation with landslide susceptibility in the training data. Notably, these western regions are underrepresented in the training dataset, and their environmental and geological distributions differ significantly from those in the data-rich eastern regions. Consequently, the model's learning has skewed toward leveraging slope as a primary predictor without adequate contextual adaptation to the unique characteristics of the less represented areas, leading to overprediction. Moreover, this overprediction of landslide susceptibility for the western area contradicts our domain knowledge. The CSF and AZ ecoregions, known for their rocky terrain (i.e., shallow soil with exposed rock), are typically associated with low shallow landslide risk on soil-mantled landscapes. The physics-based model, on the other hand, more accurately reflects this domain knowledge (comparing the landslide susceptibility of the western area in Figs. 7 and 12), using elevation-dependent sliding layer thickness (see Eq. (4)) to generate input parameters for the infinite slope model to calculate FoS values. It should be noted that the alpine environments in the Colorado Front Range are susceptible to debris flows; however, they are primarily nurtured by erosive processes (e.g., [13, 25, 75]), a triggering mechanism that is different from the landslide inventory used in the present study.

Based on the results discussed in this section (i.e., Figs. 10, 11 and 12 and Table 4), it is evident that data-driven ML models for LSM may generate predictions that do not align with our domain knowledge and generalize poorly to different ecoregions, hindering the scaling up of pure data-driven ML models.

12 Performance of PGML model

12.1 Effect of physics-guided loss function

In this study, λ_{phy} in Eq. (14) is a critical parameter that controls the influence of the physics-based loss term in regularizing PGML models. Figure 13 illustrates the relationship between PI and classification scores obtained through the spatial CV procedure, highlighting the impact of λ_{phy} on the efficacy of the physics-guided loss function for PGML models. Note that when $\lambda_{\text{phy}} = 0.0$, the PGML model is equivalent to the pure data-driven MLPNN model and a larger λ_{phy} value imposes a more stringent regularization from the physics-guided loss term. As shown in Fig. 13, an increase in λ_{phy} leads to a marked reduction in PI scores. This observed trend underscores the effectiveness of the physics-based loss term in guiding model predictions to follow the expected monotonic relationship where a higher calculated FoS value is associated with a lower landslide susceptibility. The model classification performance (i.e., Accuracy, F_1, and AUC scores) also improves with λ_{phy}, peaking at an optimum λ_{phy} value of 0.5. Beyond $\lambda_{\text{phy}} = 0.5$, the influence of the physics-based loss term becomes less pronounced, and the performance starts resembling that of the physics-based model (i.e., approaching the average values in Table 3). This observed effect can be attributed to the fact that the rule imposed by the physics-based loss term, while generally beneficial, may conflict with the actual class labels. For instance, some locations labeled as debris flows may exhibit high FoS values, which contradicts the constraints imposed by the model. Therefore, finding an appropriate trade-off between prediction accuracy and physics consistency is essential.

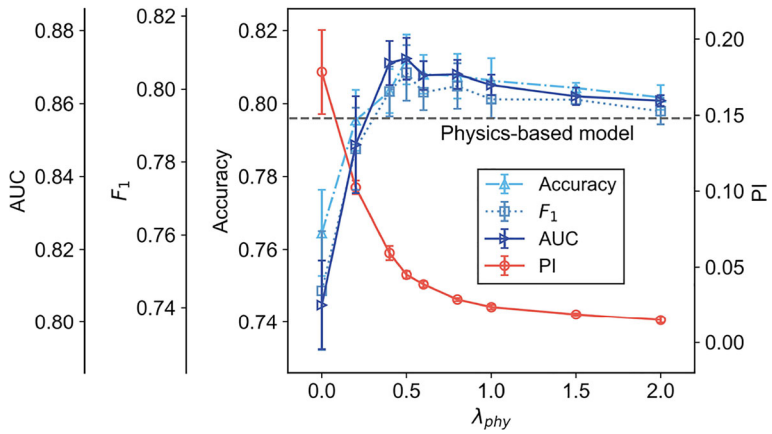

Fig. 13 Effect of physics-based loss function on PGML model performance

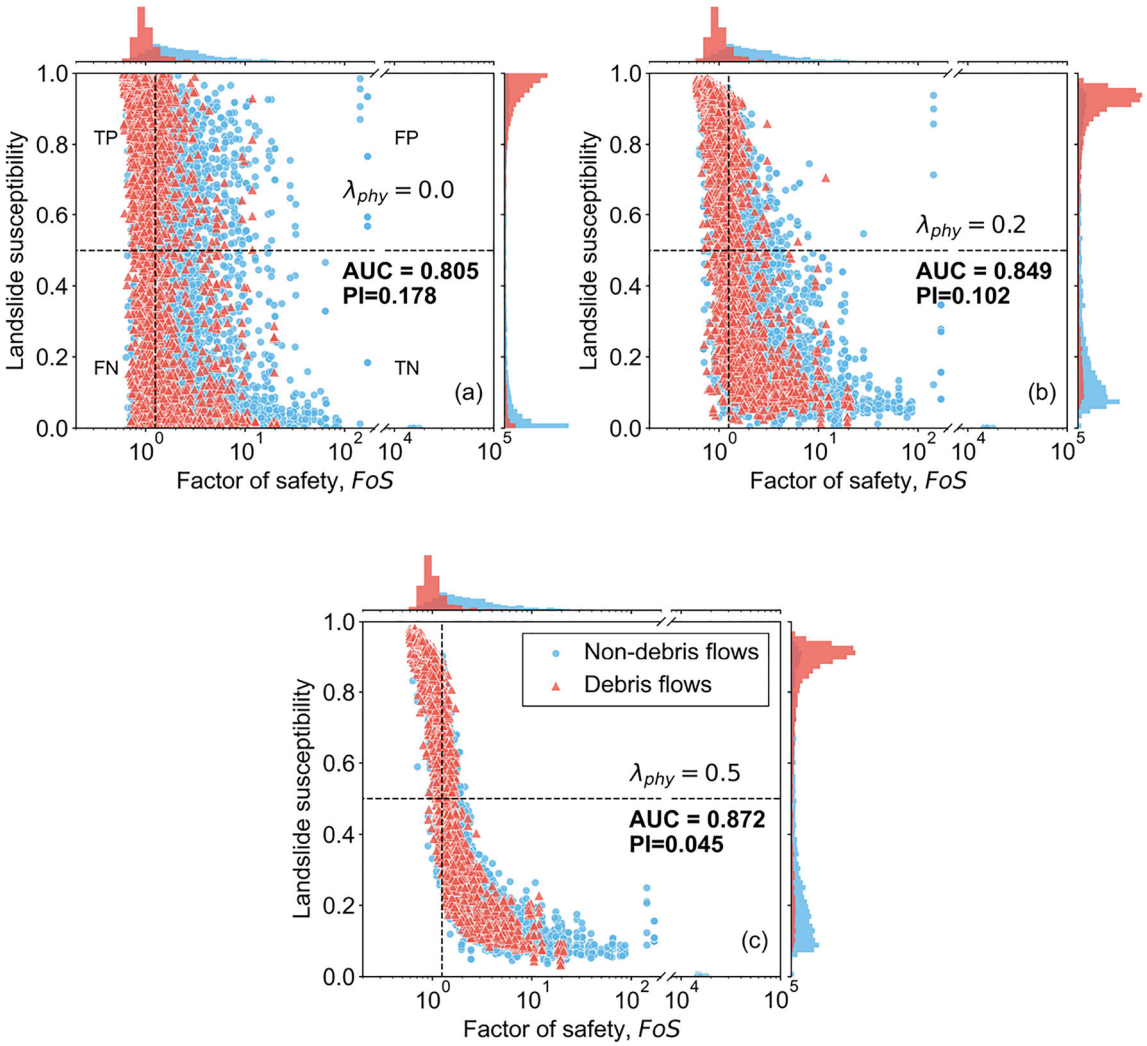

Fig. 14 Effect of physics-based loss function on enforcing physics consistency: **a** $\lambda_{phy} = 0.0$; **b** $\lambda_{phy} = 0.2$; and **c** $\lambda_{phy} = 0.5$

Figure 14. presents scatter plots with marginal histograms showing the predicted landslide susceptibility based on the PGML model versus calculated FoS using the physics-based model for validation ecoregions. To facilitate explanation, each subplot in Fig. 14 is divided into four quadrants to demonstrate the physics consistency of model predictions based on a default classification threshold of 0.5 for the PGML model and 1.25 for the physics-based model, respectively. These quadrants are: true positive (TP, landslide susceptibility > 0.5 and FoS < 1.25), true negative (TN, landslide susceptibility < 0.5 and FoS > 1.25), false positive (FP, landslide susceptibility > 0.5 and FoS > 1.25), and false negative (FN, landslide susceptibility < 0.5 and FoS < 1.25). As shown in Fig. 14a, at $\lambda_{\text{phy}} = 0.0$, the model without any constraints can distinguish debris flow and non-debris flow samples. However, the model predictions demonstrate significant physics inconsistency, as a clear monotonic relationship between the model predictions and FoS based on the physics-based model is difficult to observe. This lack of monotonicity violates our domain knowledge and indicates a misalignment between the ML model's predictions and the physics-based expectations. For example, the model incorrectly predicts low landslide susceptibilities for a substantial number of debris flow samples in the validation ecoregion that exhibit low FoS values (i.e., FN predictions). Conversely, it assigns high landslide susceptibilities to many non-debris flow samples in regions with comparatively high FoS values (i.e., FP predictions). By comparing Fig. 14a with Fig. 14b and c, it is evident that increasing the weight of λ_{phy} can significantly enhance the monotonicity of model predictions relative to the calculated FoS, making the PGML model behavior align better with our domain knowledge and reducing FP and FN predictions with respect to the physics-based model. The effects of λ_{phy} demonstrated in Figs. 13 and 14 suggest that the PGML model can harness the complementary strength

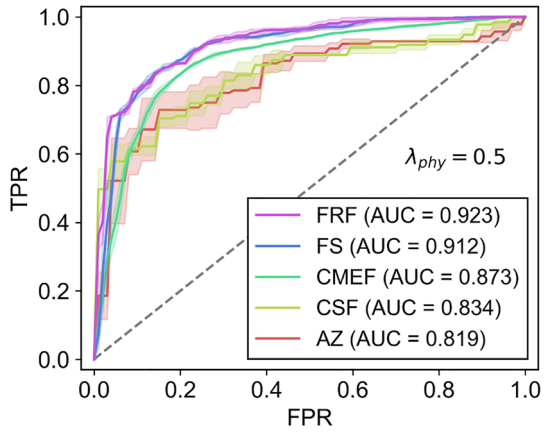

Fig. 15 ROC curve for the PGML model based on spatial CV

of the ML and the physics-based models. This allows the PGML model to excel in classification performance while maintaining high physics consistency. In addition, the discrepancies between ML and physics-based models shown in Fig. 14 highlight a critical issue: data-driven models can produce predictions that contradict established domain knowledge. Therefore, it is essential to incorporate an additional dimension, such as physics consistency metrics, to evaluate model performance and guide the training process effectively.

12.2 Performance of PGML model with optimized λ_{phy}

Figure 15 presents ROC curves for the PGML model ($\lambda_{phy} = 0.5$) for each validation ecoregion based on the spatial CV procedure with five repetitions. Lines in Fig. 15 represent mean values, while shaded areas represent standard deviations. As shown in Fig. 15, the PGML model effectively distinguishes between debris flow and non-debris flow classes within these ecoregions, with an average AUC score of 0.872. In addition, the PGML model outperforms both the physics-based model and the MLPNN model (refer to Figs. 8 and 10) in terms of classification performance. It also exhibits less fluctuation across spatial CV repetitions and demonstrates enhanced stability across different ecoregions.

Detailed classification scores can be found in Table 5, which also includes PI scores, optimal thresholds, and results based on the random CV procedure. By comparing Table 5 with Table 4, it can be noted that the PGML model showed significantly higher classification performance than the MLPNN model based on spatial CV, albeit a slight decrease in classification performance in terms of random CV. However, regardless of different CV procedures, the PGML model consistently presents a significant reduction in PI scores compared to the MLPNN model. The performance drop observed from the random CV procedure can be attributed to the physics-based loss term, which acts as a regularization factor. This term limits the model's capacity to fit the dataset by enforcing adherence to simple, domain knowledge rules. This also implies that the random CV procedure may yield misleading results when it comes to geospatial analysis, such as LSM, where the assessment of the generalization capability of the model is essential. Moreover, Table 5 also shows the optimal threshold values for the PGML model in various validation ecoregions, which are consistently closer to 0.5 compared to those for the MLPNN model in Table 4. This proximity to the conventional threshold of 0.5 also indicates the PGML model's ability to generalize more effectively across diverse ecoregions. The results in Fig. 15 and Table 5 suggest that the physics-based loss term performs effectively in steering the model toward physics consistency and reduces uncertainty in model predictions, which makes the PGML model more generalizable and robust.

Similar to the MLPNN ensemble model, the PGML ensemble model was created using the same average ensembling approach by five candidate models produced from one spatial CV repetition. This PGML ensemble model was then used to generate the landslide susceptibility map for the study area as shown in Fig. 16, which compares the landslide susceptibility map for the study area produced by the MLPNN ensemble and the PGML ensemble model. In addition, Fig. 16 also presents the

Table 5 Performance of the PGML model ($\lambda_{phy} = 0.5$) based on the cross-validation procedure

Ecoregion	Accuracy	Precision	Recall	F_1	AUC	PI	Optimal classification threshold
FRF	0.847	0.861	0.828	0.844	0.923	0.038	0.160
FS	0.849	0.856	0.840	0.847	0.912	0.049	0.517
CMEF	0.819	0.821	0.817	0.819	0.873	0.054	0.764
CSF	0.763	0.791	0.719	0.752	0.834	0.044	0.201
AZ	0.782	0.843	0.693	0.760	0.819	0.039	0.174
Avg	0.812	0.834	0.779	0.804	0.872	0.045	0.363
Random CV	0.855	0.870	0.837	0.853	0.905	0.042	0.504

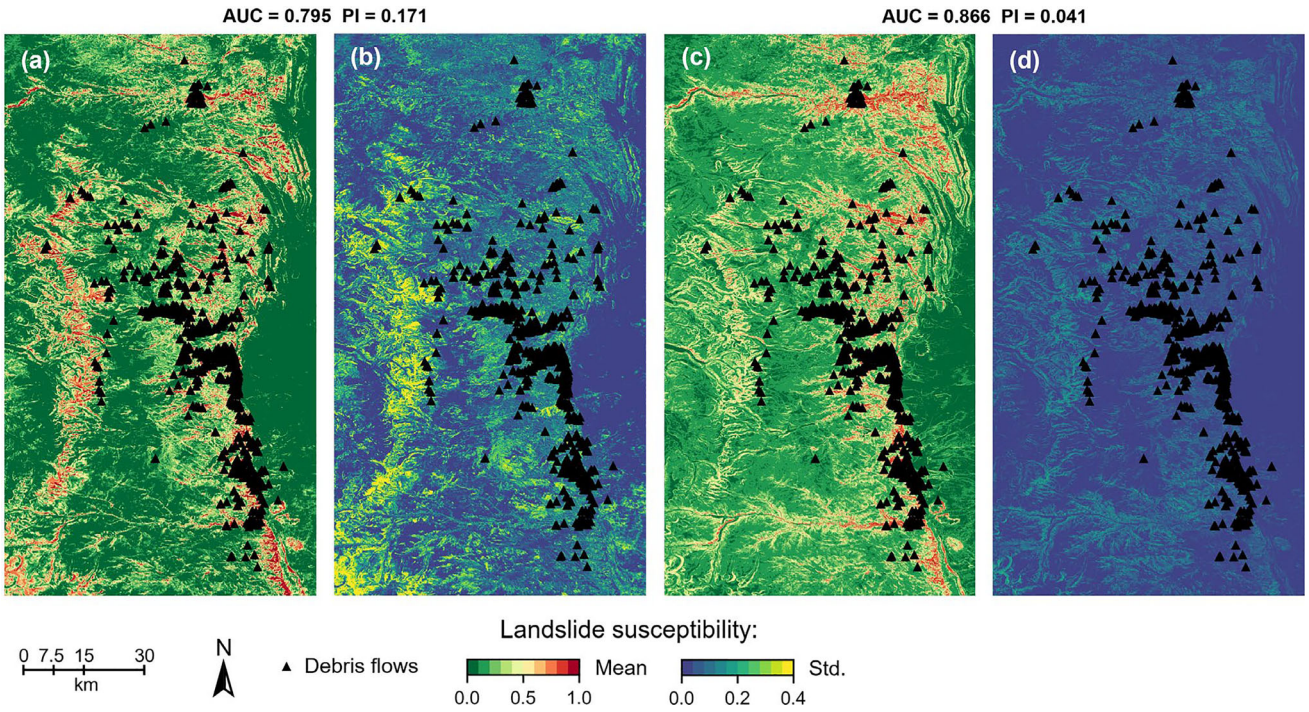

Fig. 16 Comparison of landslide susceptibility predicted by the MLPNN model and the PGML model: **a** mean value of MLPNN ensemble model outputs; **b** std. of MLPNN ensemble model outputs; **c** mean value of PGML ensemble model outputs; and **d** std. value of PGML ensemble model outputs ($\lambda_{phy} = 0.5$ was used for the PGML model)

standard deviation of model predictions for both the PGML and the MLPNN ensemble models. As shown in Fig. 16a and b, the MLPNN model overestimated landslide susceptibility for the western part of the study area encompassing the CSF and AZ ecoregions, and its predictions also reveal substantial uncertainties for these two ecoregions (note the high standard deviation in Fig. 16b). This raises concerns about the reliance on pure data-driven models (e.g., the MLPNN model) for predicting landslide susceptibility in varied environments with limited landslide inventory. In contrast, Fig. 16c and d shows that the PGML model considerably reduced overpredictions in the western part of the study area and significantly reduced prediction uncertainties across the entire study area (note the low standard deviation in Fig. 16d), which highlights the value of incorporating geotechnical domain knowledge via the physics-guide loss function in facilitating regional-scale LSM using PGML.

12.3 Generalization capability to heterogeneous ecoregions

Generalization capability is essential for LSM using ML models. This section adopted a modified spatial CV procedure to further assess the generalization capability of the PGML model across diverse ecoregions. Unlike the spatial CV procedure illustrated in Fig. 6, the training phase for the modified spatial CV procedure was executed solely on one ecoregion at a time, and the model validation was independently performed on the remaining four ecoregions. This configuration allows each model trained on a single ecoregion to be rigorously tested in other ecoregions with different environments, thereby providing a more comprehensive analysis of its generalization capabilities. It should be noted that this modified spatial CV procedure is only used in this section for model performance evaluation and is not suitable for generating landslide susceptibility maps. Using the modified spatial CV procedure, Fig. 17 compares the generalization performance between MLPNN and PGML models trained solely on the FRF ecoregion. For validation within the FRF ecoregion, the performance was evaluated using a fivefold stratified random CV, which should adequately evaluate the model's performance as data distribution within a single ecoregion is relatively uniform. For validation across different ecoregions, the outputs from these five models were aggregated into an average ensemble model. This ensemble model was subsequently utilized to assess the generalization capability across the remaining four ecoregions. As shown in Fig. 17, while both MLPNN and PGML models perform optimally in their training ecoregion (i.e., FRF), their performance deteriorates in other ecoregions, particularly as validation ecoregions become increasingly distant from the training ecoregion (i.e., the ecoregions are listed in ascending order

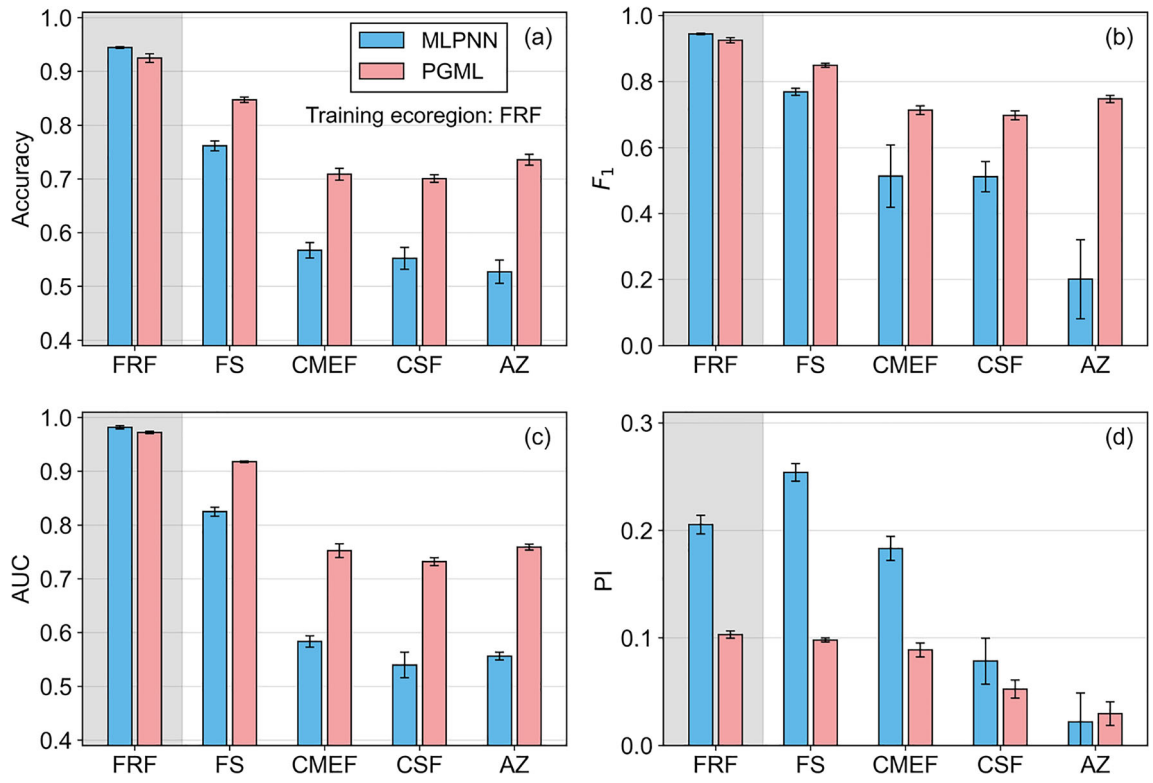

Fig. 17 Comparison of generalization performance between MLPNN and PGML models across different ecoregions: **a** accuracy; **b** F_1; **c** AUC; and **d** PI ($\lambda_{phy} = 0.5$ was used for the PGML model)

Table 6 Summary of generalization performance of ML models across different ecoregions ($\lambda_{phy} = 0.5$ was used for the PGML model)

Ecoregions	Accuracy			F_1			AUC			PI		
	MLPNN	PGML	Diff	MLPNN	PGML	Diff	MLPNN	PGML	Diff	MLPNN	PGML	Diff
FRF	0.811	0.838	3.4%	0.807	0.834	3.4%	0.856	0.884	3.3%	0.098	0.051	− 47.8%
FS	0.767	0.819	6.7%	0.772	0.821	6.4%	0.815	0.874	7.2%	0.160	0.073	− 54.5%
CMEF	0.660	0.754	14.3%	0.639	0.754	18.0%	0.700	0.806	15.2%	0.200	0.104	− 48.1%
CSF	0.678	0.752	11.0%	0.654	0.744	13.7%	0.700	0.813	16.1%	0.138	0.081	− 41.5%
AZ	0.621	0.733	18.1%	0.526	0.727	38.4%	0.638	0.767	20.2%	0.094	0.080	− 15.4%
Avg	0.707	0.779	10.7%	0.679	0.776	16.0%	0.742	0.829	12.4%	0.138	0.078	− 41.5%

in terms of elevation range in Fig. 17). However, the PGML model showed significantly less performance deterioration in all the validation ecoregions than the MLPNN models and consistently exhibited significantly low PI scores. Table 6 compares the model generalization performance between the MLPNN and the PGML models based on the modified spatial CV procedure. As shown in Table 6, the PGML model showed substantial improvements in classification performance and physics consistency compared to the MLPNN model. Based on the results presented in this section, it can be concluded that the PGML model with integrated domain knowledge can improve performance in landslide susceptibility prediction compared to the MLPNN model with better generalization capabilities across diverse ecoregions. This is validated by both the spatial CV and modified spatial CV procedures, in which less fluctuation, fewer overpredictions, and reduced prediction uncertainties can be observed for the PGML model compared to the MLPNN model.

13 Discussion and future work

This study confronts two prevalent challenges in LSM: the scarcity of data and variable conditions across diverse hillslope environments, alongside the tendency of flexible ML models to yield predictions that deviate from established domain knowledge in underrepresented areas. The

proposed PGML framework integrates geotechnical domain knowledge into ML paradigms, ensuring predictions are both empirically grounded and theoretically consistent.

A key innovation of this approach is the introduction of performance metrics based on physics consistency, which complements the commonly used data science model evaluation metrics and helps quantify the extent to which model predictions adhere to physical principles, providing an essential measure of reliability in geoscientific applications. Additionally, a cross-validation strategy that accounts for the inherent structural dependencies within the data was employed, demonstrating its effectiveness in enhancing model performance assessment at a regional scale. The performance of the proposed PGML framework was evaluated through a case study in the Colorado Front Range, employing a well-documented debris flow inventory to compare its effectiveness with both a physics-based infinite slope model and a purely data-driven MLPNN model.

The proposed PGML framework is not only pivotal for LSM but also offers a generic solution applicable to various challenges within geotechnical engineering and geoscience. For example, it can be extended to issues involving data with temporal, spatial, hierarchical, or phylogenetic structures, such as hydrological modeling, soil erosion modeling, etc. These applications, like LSM, require predictions that align closely with domain knowledge to ensure the validity and applicability of outcomes. By enhancing the integration and utility of domain-specific knowledge, the proposed framework can substantially improve the accuracy and generalizability of predictive models across extensive geoscientific domains.

It should be noted that the present study conducted the physics-based analysis using detailed landslide reconnaissance and previous studies that reveal the general environmental conditions for the study area. Such information may not be readily available for LSM in other regions. Therefore, exploring various strategies to incorporate geotechnical domain knowledge into ML models and examining the effects of sample size warrant further investigation in future research. For example, innovative approaches such as differentiable modeling [83], which aims to unify physics and ML by embedding learnable parameters within process-based modeling, offer promising pathways for enhancing LSM frameworks. Furthermore, the use of unsupervised pretext task learning [35, 89] might be a viable approach to develop foundation models for geospatial features, which can then be used for downstream tasks in landslide hazard modeling.

14 Conclusion

The findings from this study underscore the effectiveness of the proposed PGML framework in enhancing model performance and reliability. Key conclusions drawn from this study can be summarized as follows:

1. The random CV approach may produce overly optimistic and sometimes misleading results that may not reflect the actual generalization performance of the model, and spatial CV is a more suitable approach for geospatial applications.
2. By using existing domain knowledge in geotechnical engineering to identify appropriate input parameters and models, the physics-based infinite slope model can effectively predict regional-scale landslide susceptibility.
3. The pure data-driven model (i.e., MLPNN) model generally performs poorly on unseen ecoregions and exhibits significant uncertainties in model predictions for regions lacking sufficient debris flow inventory.
4. By integrating geotechnical domain knowledge into pure data-driven models, the PGML model exhibits significant improvements in generalization performance, better physics consistency, and reduced uncertainties.

Acknowledgements This research was partially supported by Google AI Impacts Challenge Grant 1904-57775. The second author's effort was also supported by the U.S. National Science Foundation under Award No. ICER-2022444. This support is gratefully acknowledged. The authors would also like to acknowledge Dr. Chaopeng Shen and Dr. Daniel Kifer from Pennsylvania State University for their insightful input during the planning and development of this work and two anonymous reviewers for their valuable comments and suggestions, which helped us improve the quality of the manuscript.

Author contribution TP contributed to conceptualization, methodology, data curation, software, investigation, and writing—original draft; TQ contributed to conceptualization, writing—review & editing, supervision, and funding acquisition. All authors reviewed the results and approved the final version of the manuscript.

Data availability Some data, models, or codes that support the findings of this study are available from the corresponding author upon reasonable request. These data include the results used to generate all figures and tables.

Declarations

Competing interests The authors declare no competing interests.

References

1. Anderson RS, Riihimaki CA, Safran EB, MacGregor KR (2006) Facing reality: late Cenozoic evolution of smooth peaks, glacially ornamented valleys, and deep river gorges of Colorado's

Front Range. In: Tectonics, climate, and landscape evolution. Geological Society of America, pp 397–418. https://doi.org/10.1130/2006.2398(25)
2. Anderson SW, Anderson SP, Anderson RS (2015) Exhumation by debris flows in the 2013 Colorado Front Range storm. Geology 43:391–394. https://doi.org/10.1130/g36507.1
3. Agarap AF (2018) Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375. https://doi.org/10.48550/arXiv.1803.08375
4. Bowles JE (1979) Physical and geotechnical properties of soils. McGraw-Hill, Incorporated, New York, NY
5. Birkeland PW, Shroba RR, Burns SF, Price AB, Tonkin PJ (2003) Integrating soils and geomorphology in mountains—an example from the Front Range of Colorado. Geomorphology 55:329–344. https://doi.org/10.1016/s0169-555x(03)00148-x
6. Brenning A (2005) Spatial prediction models for landslide hazards: review, comparison and evaluation. Nat Hazard 5:853–862. https://doi.org/10.5194/nhess-5-853-2005
7. Baum RL, Godt JW, Savage WZ (2010) Estimating the timing and location of shallow rainfall-induced landslides using a model for transient, unsaturated infiltration. J Geophys Res. https://doi.org/10.1029/2009jf001321
8. 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. https://doi.org/10.1109/igarss.2012.6352393
9. Baum RL, Scheevel CR, Jones ES (2019) Constraining parameter uncertainty in modeling debris-flow initiation during the september 2013 Colorado Front Range storm. In: Association of environmental and engineering geologists; special publication 28. Colorado School of Mines. Arthur Lakes Library. https://doi.org/10.25676/11124/173212
10. Cox DR (1959) The regression analysis of binary sequences. J R Stat Soc 21:238–238. https://doi.org/10.1111/j.2517-6161.1959.tb00334.x
11. Cruden DM, Varnes DJ (1996) Landslide types and processes. Transportation Research Board, U.S. National Academy of Sciences special report. Transp Res Board 247(1996):36–57
12. Castellanos Abella EA, Van Westen CJ (2008) Qualitative landslide susceptibility assessment by multicriteria analysis: a case study from San Antonio del Sur, Guantánamo. Cuba Geomorphol (Amst) 94:453–466. https://doi.org/10.1016/j.geomorph.2006.10.038
13. Coe JA, Kinner DA, Godt JW (2008) Initiation conditions for debris flows generated by runoff at Chalk Cliffs, central Colorado. Geomorphology (Amst) 96:270–297. https://doi.org/10.1016/j.geomorph.2007.03.017
14. Coe JA, Kean JW, Godt JW, Baum RL, Jones ES, Gochis DJ, Anderson GS (2014) New insights into debris-flow hazards from an extraordinary event in the Colorado Front Range. GSA Today 24:4–10. https://doi.org/10.1130/gsatg214a.1
15. Carvalho DV, Pereira EM, Cardoso JS (2019) Machine learning interpretability: a survey on methods and metrics. Electronics 8:832. https://doi.org/10.3390/electronics8080832
16. Chen L, Ding Y, Pirasteh S, Hu H, Zhu Q, Ge X, Zeng H, Yu H, Shang Q, Song Y (2022) Meta-learning an intermediate representation for few-shot prediction of landslide susceptibility in large areas. Int J Appl Earth Obs Geoinf 110:102807. https://doi.org/10.1016/j.jag.2022.102807
17. Dickinson WR, Klute MA, Hayes MJ, Janecke SU, Lundin MA, McKittrick MA, Olivares MD (1988) Paleogeographic and paleotectonic setting of Laramide sedimentary basins in the central Rocky Mountain region. Geol Soc Am Bull 100:1023–1039. https://doi.org/10.1130/0016-7606(1988)100%3c1023:papsol%3e2.3.co;2
18. Daw A, Karpatne A, Watkins W, Read J, Kumar V (2017) Physics-guided neural networks (PGNN): an application in lake temperature modeling." arXiv [cs.LG]. https://doi.org/10.48550/arXiv.1710.11431
19. Dong X, Yu Z, Cao W, Shi Y, Ma Q (2020) A survey on ensemble learning. Front Comput Sci 14(2):241–258. https://doi.org/10.1007/s11704-019-8208-z
20. Di Napoli M, Carotenuto F, Cevasco A, Confuorto P, Di Martire D, Firpo M, Pepe G, Raso E, Calcaterra D (2020) Machine learning ensemble modelling as a tool to improve landslide susceptibility mapping reliability. Landslides 17(8):1897–1914. https://doi.org/10.1007/s10346-020-01392-9
21. Fischer E, Knutti R (2015) Anthropogenic contribution to global occurrence of heavy-precipitation and high-temperature extremes. Nat Clim Chang 5:560–564. https://doi.org/10.1038/nclimate2617
22. Froude MJ, Petley DN (2018) Global fatal landslide occurrence from 2004 to 2016. Nat Hazard 18:2161–2181. https://doi.org/10.5194/nhess-18-2161-2018
23. Fang Z, Wang Y, Peng L, Hong H (2021) A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility mapping. Geogr Inf Syst 35(2):321–347. https://doi.org/10.1080/13658816.2020.1808897
24. Guzzetti F, Reichenbach P, Cardinali M, Galli M, Ardizzone F (2005) Probabilistic landslide hazard assessment at the basin scale. Geomorphology 72(1–4):272–299. https://doi.org/10.1016/j.geomorph.2005.06.002
25. Godt JW, Coe JA (2007) Alpine debris flows triggered by a 28 July 1999 thunderstorm in the central Front Range, Colorado. Geomorphology (Amst) 84:80–97. https://doi.org/10.1016/j.geomorph.2006.07.009
26. Grover A, Kapoor A, Horvitz E (2015) A deep hybrid model for weather forecasting. In: Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA. https://doi.org/10.1145/2783258.2783275
27. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. The MIT Press, Cambridge, MA
28. Gupta M, Cotter A, Pfeifer J, Voevodski K, Canini K, Mangylov A, Moczydlowski W, van Esbroeck A (2016) Monotonic calibrated interpolated look-up tables. J Mach Learn Res 17(109):1–47
29. Johnston EC, Davenport FV, Wang L, Caers JK, Muthukrishnan S, Burke M, Diffenbaugh NS (2021) Quantifying the effect of precipitation on landslide hazard in urbanized and non-urbanized areas. Geophys Res Lett 48:e2021GL094038. https://doi.org/10.1029/2021gl094038
30. Jia X, Willard J, Karpatne A, Read JS, Zwart JA, Steinbach M, Kumar V (2021) Physics-guided machine learning for scientific discovery: an application in simulating lake temperature profiles. ACM/IMS Trans Data Sci ACM/IMS Trans Data Sci 2(3):1–26. https://doi.org/10.1145/3447814
31. Huang J-C, Kao S-J, Hsu M-L, Lin J-C (2006) Stochastic procedure to extract and to integrate landslide susceptibility maps: an example of mountainous watershed in Taiwan. Nat Hazards Earth Syst Sci 6:803–815. https://doi.org/10.5194/nhess-6-803-2006
32. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York, NY
33. He X, Hong Y, Vergara H, Zhang K, Kirstetter PE, Gourley JJ, Zhang Y, Qiao G, Liu C (2016) Development of a coupled hydrological-geotechnical framework for rainfall-induced landslides prediction. J Hydrol (Amst) 543:395–405. https://doi.org/10.1016/j.jhydrol.2016.10.016

34. Hengl T, Mendes de Jesus J, Heuvelink GBM, Ruiperez Gonzalez M, Kilibarda M, Blagotić A et al (2017) SoilGrids250m: global gridded soil information based on machine learning. PLoS ONE 12:e0169748. https://doi.org/10.1371/journal.pone.0169748
35. Hong D, Li C, Zhang B, Yokoya N, Benediktsson JA, Chanussot J (2024) Multimodal artificial intelligence foundation models: unleashing the power of remote sensing big data in earth observation. Innov Geosci 2(1):100055. https://doi.org/10.59717/j.xinn-geo.2024.100055
36. Kirschbaum DB, Adler R, Hong Y, Hill S, Lerner-Lam A (2010) A global landslide catalog for hazard applications: method, results, and limitations. Nat Hazards 52(3):561–575. https://doi.org/10.1007/s11069-009-9401-4
37. Kuhn M, Johnson K (2013) Applied predictive modeling. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-6849-3
38. Karpatne A, Atluri G, Faghmous JH, Steinbach M, Banerjee A, Ganguly A, Shekhar S, Samatova N, Kumar V (2017) Theory-guided data science: a new paradigm for scientific discovery from data. IEEE Trans Knowl Data Eng 29:2318–2331. https://doi.org/10.1109/tkde.2017.2720168
39. Khabiri S, Crawford MM, Koch HJ, Haneberg WC, Zhu Y (2023) An assessment of negative samples and model structures in landslide susceptibility characterization based on Bayesian network models. Remote Sens (Basel) 15(12):3200. https://doi.org/10.3390/rs15123200
40. Lee JH, Park HJ (2016) Assessment of shallow landslide susceptibility using the transient infiltration flow model and GIS-based probabilistic approach. Landslides 13:885–903. https://doi.org/10.1007/s10346-015-0646-6
41. Lombardo L, Opitz T, Ardizzone F, Guzzetti F, Huser R (2020) Space-time landslide predictive modelling. Earth Sci Rev 209:103318. https://doi.org/10.1016/j.earscirev.2020.103318
42. Lv L, Chen T, Dou J, Plaza A (2022) A hybrid ensemble-based deep-learning framework for landslide susceptibility mapping. Int J Appl Earth Obs Geoinf 108(102713):102713. https://doi.org/10.1016/j.jag.2022.102713
43. 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
44. Liu S, Wang L, Zhang W, Sun W, Fu J, Xiao T, Dai Z (2023) A physics-informed data-driven model for landslide susceptibility assessment in the three Gorges Reservoir area. Geosci Front 14(5):101621. https://doi.org/10.1016/j.gsf.2023.101621
45. Li Z, Pei T, Ying W, Srubar WV III, Zhang R, Yoon J, Ye H, Dabo I, Radlińska A (2024) Can domain knowledge benefit machine learning for concrete property prediction? J Am Ceram Soc 107(3):1582–1602. https://doi.org/10.1111/jace.19549
46. Li Z, Pei T, Ying W, Srubar WV III, Zhang R, Yoon J, Ye H, Dabo I, Radlińska A (2024) Simulation-based transfer learning for concrete strength prediction. In: Banthia N, Soleimani-Dashtaki S, Mindess S (eds) Smart & sustainable infrastructure: building a greener tomorrow. ISSSI 2023. RILEM Bookseries, vol 48. Springer, Cham. https://doi.org/10.1007/978-3-031-53389-1_98
47. Marr JW (1961) Ecosystems of the east slope of the Front Range in Colorado. University of Colorado Studies, Series in Biology Number 8. University of Colorado Press: Boulder, CO, USA
48. Mcmahon E, Gregonis SM, Waltman SW, Omernik JM, Thorson TD, Freeouf JA, Rorick AH, Keys JE (2001) Developing a spatial framework of common ecological regions for the conterminous United States. Environ Manag 28:293–316. https://doi.org/10.1007/s0026702429
49. Montrasio L, Valentino R (2008) A model for triggering mechanisms of shallow landslides. Nat Hazards Earth Syst Sci 8:1149–1159. https://doi.org/10.5194/nhess-8-1149-2008
50. McGuire LA, Rengers FK, Kean JW et al (2016) Elucidating the role of vegetation in the initiation of rainfall-induced shallow landslides: insights from an extreme rainfall event in the Colorado Front Range. Geophys Res Lett 43:9084–9092. https://doi.org/10.1002/2016GL070741
51. 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
52. Merghadi A, Yunus AP, Dou J, Whiteley J, ThaiPham B, Bui DT, 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
53. Mirus BB, Jones ES, Baum RL et al (2020) Landslides across the USA: occurrence, susceptibility, and data limitations. Landslides 17:2271–2285. https://doi.org/10.1007/s10346-020-01424-4
54. Ma K, Feng D, Lawson K, Tsai W-P, Liang C, Huang X, Sharma A, Shen C (2021) Transferring hydrologic data across continents—leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. Water Resour Res. https://doi.org/10.1029/2020WR028600
55. Medina V, Hürlimann M, Guo Z, Lloret A, Vaunat J (2021) Fast physically-based model for rainfall-induced landslide susceptibility assessment at regional scale. CATENA 201:105213. https://doi.org/10.1016/j.catena.2021.105213
56. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2022) A survey on bias and fairness in machine learning. ACM Comput Surv 54(6):1–35. https://doi.org/10.1145/3457607
57. Montrasio L, Gatto MP, Miodini C (2023) The role of plants in the prevention of soil-slip: the G-SLIP model and its application on territorial scale through G-XSLIP platform. Landslides 20:1149–1165. https://doi.org/10.1007/s10346-023-02031-9
58. Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal M-E, Ruggieri S, Turini F, Papadopoulos S, Krasanakis E, Kompatsiaris I, Kinder-Kurlanda K, Wagner C, Karimi F, Fernandez M, Alani H, Berendt B, Kruegel T, Heinze C, Broelemann K, Kasneci G, Tiropanis T, Staab S (2020) Bias in data-driven artificial intelligence systems—an introductory survey. Wiley Interdiscip Rev Data Min Knowl Discov. https://doi.org/10.1002/widm.1356
59. Nagendra S, Kifer D, Mirus B, Pei T, Lawson K, Manjunatha SB, Li W, Nguyen H, Qiu T, Tran S, Shen C (2022) Constructing a large-scale landslide database across heterogeneous environments using task-specific model updates. IEEE J Sel Top Appl Earth Obs Remote Sens 15:4349–4370. https://doi.org/10.1109/jstars.2022.3177025
60. Pellicani R, Frattini P, Spilotro G (2014) Landslide susceptibility assessment in Apulian Southern Apennine: heuristic versus statistical methods. Environ Earth Sci 72:1097–1108. https://doi.org/10.1007/s12665-013-3026-3
61. Piciullo L, Calvello M, Cepeda JM (2018) Territorial early warning systems for rainfall-induced landslides. Earth Sci Rev 179:228–247. https://doi.org/10.1016/j.earscirev.2018.02.013
62. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) PyTorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 32:8024–8035

63. Pawar S, San O, Aksoylu B, Rasheed A, Kvamsdal T (2021) Physics guided machine learning using simplified theories. Phys Fluids 33(1):011701. https://doi.org/10.1063/5.0038929
64. Pei T, Nagendra S, Banagere Manjunatha S, He G, Kifer D, Qiu T, Shen C (2021) Utilizing an interactive AI-empowered web portal for landslide labeling for establishing a landslide database in Washington state, USA, EGU General Assembly 2021, online, 19–30 Apr 2021, EGU21-13974. https://doi.org/10.5194/egusphere-egu21-13974
65. Pei T, Qiu T (2023) Debris flow susceptibility mapping in Colorado Front Range, USA: A comparison of physics-based and data-driven approaches. E3S Web of Conferences 415:01018. https://doi.org/10.1051/e3sconf/202341501018
66. Pei T, Qiu T, Shen C (2023) Applying knowledge-guided machine learning to slope stability prediction. J Geotech Geoenviron Eng 149(10):04023089. https://doi.org/10.1061/jggefk.gteng-11053
67. Pei T, Qiu T (2023) Machine learning with monotonic constraint for geotechnical engineering applications: an example of slope stability prediction. Acta Geotech. https://doi.org/10.1007/s11440-023-02117-7
68. Pei T, Qiu T (2023) Landslide susceptibility mapping using machine learning methods: a case study in Colorado front range, USA. In: Geo-Congress 2023. American Society of Civil Engineers, Reston, VA. https://doi.org/10.1061/9780784484654.052
69. Pei T, Liu J, Shen C, Kifer D (2023) Impact of cross-validation strategies on machine learning models for landslide susceptibility mapping: a comparative study. AGU Fall Meeting Abstracts 2023, NH13D-0717
70. Ruff M, Czurda K (2008) Landslide susceptibility analysis with a heuristic approach in the Eastern Alps (Vorarlberg, Austria). Geomorphology (Amst) 94:314–324. https://doi.org/10.1016/j.geomorph.2006.10.032
71. Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI, Wintle BA, Hartig F, Dormann CF (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40:913–929. https://doi.org/10.1111/ecog.02881
72. 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
73. Read JS, Jia X, Willard J, Appling AP, Zwart JA, Oliver SK, Karpatne A, Hansen GJA, Hanson PC, Watkins W, Steinbach M, Kumar V (2019) Process-guided deep learning predictions of lake water temperature. Water Resour Res 55(11):9173–9190. https://doi.org/10.1029/2019WR024922
74. Rai R, Sahu CK (2020) Driven by data or derived through physics? A review of hybrid physics guided machine learning techniques with cyber-physical system (CPS) focus. IEEE Access 8:71050–71073. https://doi.org/10.1109/ACCESS.2020.2987324
75. Rengers FK, Kean JW, Reitman NG et al (2020) The influence of frost weathering on debris flow sediment supply in an alpine basin. J Geophys Res Earth Surf 125:e2019JF005369. https://doi.org/10.1029/2019jf005369
76. Rahmani F, Appling A, Feng D, Lawson K, Shen C (2023) Identifying structural priors in a hybrid differentiable model for stream water temperature modeling. Water Resour Res. https://doi.org/10.1029/2023wr034420
77. Saulnier G-M, Beven K, Obled C (1997) Including spatially variable effective soil depths in TOPMODEL. J Hydrol (Amst) 202:158–172. https://doi.org/10.1016/s0022-1694(97)00059-0
78. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manag 45:427–437. https://doi.org/10.1016/j.ipm.2009.03.002
79. Steger S, Brenning A, Bell R, Glade T (2016) The propagation of inventory-based positional errors into statistical landslide susceptibility models. Nat Hazard 16:2729–2745. https://doi.org/10.5194/nhess-16-2729-2016
80. Schratz P, Muenchow J, Iturritxa E (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
81. Stanley TA, Kirschbaum DB, Sobieszczyk S, Jasinski MF, Borak JS, Slaughter SL (2020) Building a landslide hazard indicator with machine learning and land surface models. Environ Model Softw 129:104692. https://doi.org/10.1016/j.envsoft.2020.104692
82. 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 9:640043. https://doi.org/10.3389/feart.2021.640043
83. Shen C, Appling AP, Gentine P, Bandai T, Gupta H, Tartakovsky A, Baity-Jesi M, Fenicia F, Kifer D, Li L, Liu X, Ren W, Zheng Y, Harman CJ, Clark M, Farthing M, Feng D, Kumar P, Aboelyazeed D, Rahmani F, Song Y, Beck HE, Bindas T, Dwivedi D, Fang K, Höge M, Rackauckas C, Mohanty B, Roy T, Xu C, Lawson K (2023) Differentiable modelling to unify machine learning and physical models for geosciences. Nat Rev Earth Environ 4(8):552–567. https://doi.org/10.1038/s43017-023-00450-9
84. Tiwari B, Marui H (2005) A new method for the correlation of residual shear strength of the soil with mineralogical composition. J Geotech Geoenviron Eng 131:1139–1150. https://doi.org/10.1061/(ASCE)1090-0241(2005)131:9(1139)
85. Tsai W-P, Feng D, Pan M, Beck H, Lawson K, Yang Y, Liu J, Shen C (2021) From calibration to parameter learning: harnessing the scaling effects of big data in geoscientific modeling. Nat Commun. https://doi.org/10.1038/s41467-021-26107-z
86. U.S. Geological Survey (USGS) (2017). 1/3rd arc-second digital elevation models (DEMs)—USGS National Map 3DEP downloadable data collection: U.S. Geological Survey
87. van der Maaten L, Hinton GE (2015) Visualizing high-dimensional data using t-SNE. J Mach Learn Res 9:2579–2605
88. von Rueden L, Mayer S, Beckh K, Georgiev B, Giesselbach S, Heese R, Kirsch B, Walczak M, Pfrommer J, Pick A, Ramamurthy R, Garcke J, Bauckhage C, Schuecker J (2021) Informed machine learning—a taxonomy and survey of integrating prior knowledge into learning systems. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2021.3079836
89. Vahdani E, Jing L, Huenerfauth M, Tian Y (2024) Multi-modal multi-channel American sign language recognition. Int J Artif Intell Robot Res. https://doi.org/10.1142/s2972335324500017
90. Whiteley JS, Chambers JE, Uhlemann S, Wilkinson PB, Kendall JM (2019) Geophysical monitoring of moisture-induced landslides: a review. Rev Geophys 57:106–145. https://doi.org/10.1029/2018rg000603
91. Wang S, Zhang K, van Beek LPH, Tian X, Bogaard TA (2020) Physically-based landslide prediction over a large region: scaling low-resolution hydrological model results for high-resolution slope stability assessment. Environ Model Softw 124:104607. https://doi.org/10.1016/j.envsoft.2019.104607
92. Wang Y, Yao Q, Kwok JT, Ni LM (2021) Generalizing from a few examples: a survey on few-shot learning. ACM Comput Surv 53(3):1–34. https://doi.org/10.1145/3386252
93. Wang Z, Goetz J, Brenning A (2022) Transfer learning for landslide susceptibility modeling using domain adaptation and

case-based reasoning. Geosci Model Dev 15(23):8765–8784. https://doi.org/10.5194/gmd-15-8765-2022
94. Willard J, Jia X, Xu S, Steinbach M, Kumar V (2023) Integrating scientific knowledge with machine learning for engineering and environmental systems. ACM Comput Surv 55:1–37. https://doi.org/10.1145/3514228
95. Woodard JB, Mirus BB, Crawford MM, Or D, Leshchinsky BA, Allstadt KE, Wood NJ (2023) Mapping landslide susceptibility over large regions with limited data. J Geophys Res Earth Surf. https://doi.org/10.1029/2022jf006810
96. Wang H, Wang L, Zhang L (2023) Transfer learning improves landslide susceptibility assessment. Gondwana Res 123:238–254. https://doi.org/10.1016/j.gr.2022.07.008
97. Wei X, Zhang L, Gardoni P, Chen Y, Tan L, Liu D, Du C, Li H (2023) Comparison of hybrid data-driven and physical models for landslide susceptibility mapping at regional scales. Acta Geotech 18(8):4453–4476. https://doi.org/10.1007/s11440-023-01841-4
98. Wei X, Gardoni P, Zhang L, Tan L, Liu D, Du C, Li H (2024) Improving pixel-based regional landslide susceptibility mapping. Geosci Front 15(4):101782. https://doi.org/10.1016/j.gsf.2024.101782
99. Xiong J, Pei T, Qiu T (2023) A machine learning-based method with integrated physics knowledge for predicting bearing capacity of pile foundations. In: Geo-congress 2023. American Society of Civil Engineers, Reston, VA. https://doi.org/10.1061/9780784484685.018
100. Yang T, Sun F, Gentine P, Liu W, Wang H, Yin J, Du M, Liu C (2019) Evaluation and machine learning improvement of global hydrological model-based flood simulations. Environ Res Lett 14(11):114027. https://doi.org/10.1088/1748-9326/ab4d5e
101. Zhang LL, Zhang J, Zhang LM, Tang WH (2011) Stability analysis of rainfall-induced slope failure: a review. Proc Inst Civ Eng Geotech Eng 164:299–316. https://doi.org/10.1680/geng.2011.164.5.299
102. Zhang L, Wang G, Giannakis GB (2019) Real-time power system state estimation and forecasting via deep unrolled neural networks. IEEE Trans Signal Process IEEE Trans Signal Process 67(15):4069–4077. https://doi.org/10.1109/TSP.2019.2926023
103. Zhu Q, Chen L, Hu H (2020) Unsupervised feature learning to improve transferability of landslide susceptibility representations. IEEE J Select Top Appl Earth Observ Remote Sens 13:3917–3930. https://doi.org/10.1109/jstars.2020.3006192
104. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O (2021) Understanding deep learning (still) requires rethinking generalization. Commun ACM 64(3):107–115. https://doi.org/10.1145/3446776
105. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q (2021) A comprehensive survey on transfer learning. Proc IEEE Inst Electr Electron Eng 109(1):43–76. https://doi.org/10.1109/jproc.2020.3004555
106. Zhang X, Yu W, Pun M-O, Shi W (2023) Cross-domain landslide mapping from large-scale remote sensing images using prototype-guided domain-aware progressive representation learning. ISPRS J Photogramm Remote Sens 197:1–17. https://doi.org/10.1016/j.isprsjprs.2023.01.018
107. Zeng T, Wu L, Peduto D, Glade T, Hayakawa YS, Yin K (2023) Ensemble learning framework for landslide susceptibility mapping: different basic classifier and ensemble strategy. Geosci Front 14(6):101645. https://doi.org/10.1016/j.gsf.2023.101645
108. Zhou Z, Zhang L, Zhang Q, Hu C, Wang G, She D, Chen J (2024) Global increase in future compound heat stress-heavy precipitation hazards and associated socio-ecosystem risks. NPJ Clim Atmos Sci. https://doi.org/10.1038/s41612-024-00579-4