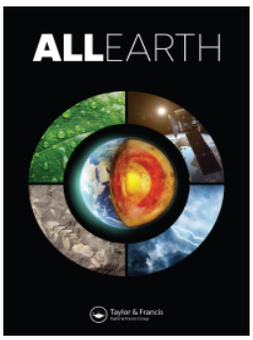

All Earth

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgda21

Proposing an easy-to-use tool for estimating landslide dimensions using a data-driven approach

Minu Treesa Abraham, Neelima Satyam, Biswajeet Pradhan & Samuele Segoni

To cite this article: Minu Treesa Abraham, Neelima Satyam, Biswajeet Pradhan & Samuele Segoni (2022) Proposing an easy-to-use tool for estimating landslide dimensions using a data-driven approach, All Earth, 34:1, 243-258, DOI: 10.1080/27669645.2022.2127549

To link to this article: https://doi.org/10.1080/27669645.2022.2127549

Published online: 27 Sep 2022.

ALL EARTH
2022, VOL. 34, NO. 1, 243–258
https://doi.org/10.1080/27669645.2022.2127549

Taylor & Francis
Taylor & Francis Group

Proposing an easy-to-use tool for estimating landslide dimensions using a data-driven approach

Minu Treesa Abraham[a], Neelima Satyam[a], Biswajeet Pradhan[b,c,d] and Samuele Segoni[e]

[a]Department of Civil Engineering, Indian Institute of Technology Indore, Indore, India; [b]Centre for Advanced Modeling and Geospatial Information Systems (CAMGIS), School of Civil and Environmental Engineering, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia; [c]Center of Excellence for Climate Change Research, King Abdulaziz University, Jeddah, Saudi Arabia; [d]Earth Observation Centre, Institute of Climate Change, Universiti Kebangsaan Malaysia, Bangi, Malaysia; [e]Department of Earth Sciences, University of Florence, Florence, Italy

ABSTRACT

The increase in population and urbanisation of hilly regions have increased the risk due to landslides. This manuscript presents a data-driven approach with a random forest algorithm to estimate the projected area, length, travel distance, and width of landslides, using elevation and slope information. The method is tested for two different study areas (Idukki and Wayanad), using three different combinations of inputs. The input features considered were elevation (*E*), tangential slope (θ), drop height (*H*), angle of reach (α) and the profile curvature (*c*). A total of 144 models were considered and were evaluated using mean-absolute-error (*MAE*) and root-mean-square-error (RMSE) values. The results indicate that, by using E and θ alone, the *RMSE* value in estimating the length values for flow-like landslides in Wayanad was reduced from 472.74 m to 204.64 m. Out of the 48 combinations considered, *MAE* values have increased in seven cases and *RMSE* values in eight cases only. The pre-trained models are saved and used to develop an easy-to-use tool, which can bypass the complications associated with the existing statistical approaches. The tool can be used by untrained personnel for preliminary hazard assessment.

ARTICLE HISTORY
Received 26 May 2022
Accepted 20 September 2022

KEYWORDS
landslides; hazard; random forest; travel distance; machine learning

1. Introduction

Landslides are common natural hazards in hilly regions, responsible for severe economic loss and casualties across the globe. With the change in climate and the increased number of extreme rainfall events, the number of rainfall-induced landslides has also increased (Gariano & Guzzetti, 2016). The local people are the first ones to identify the tension cracks or minor failures before the occurrence of a landslide. If the area that may get affected by the failure can be effectively communicated to them based on the available information, it can be helpful in making the action plan and communicating the same with the stakeholders. In a recent study, it was stated that people are more likely to trust early warnings than structural mitigation measures (Huang et al., 2021). Understanding the failure mechanism and the post-failure movement of landslides is an essential part of the assessment of a landslide hazard. The attempts to understand the mechanism of failure and the post-failure motion of slope instabilities had started in the early 20th century itself (Terzaghi, 1950). The system of landslides and their evolutions has different temporal scales, and the variations with respect to time involve different stages such as deformations before the failure, the failure itself, and the displacements after the failure. The term failure is critical, as it decides the separation of phases in the process. Failure indicates the first formation of a rupture surface as displacement (Leroueil et al., 1996), and it happens when the factor of safety (*FS*) becomes lesser than 1. This stage involves aProposing an easy-to-use tool for estimating landslide dimensions change in kinematic behaviour, from sliding to fall or flow. This change is also critical in deciding the post-failure behaviour. This analysis can be carried out either in a forensic style, as a back analysis, or as a prediction for future events.

The displacement post-failure is highly dependent on the type of failure. The knowledge of landslide typology is critical in analysing the post-failure motion, and for precise runout analysis, complex process-based models should be used separately for each landslide type (Armento et al., 2008; Guzzetti et al., 2002). Many empirical, analytical, and dynamic models have been used to quantify the post-failure motion of landslides. The advancements in numerical models have helped in understanding the triggering and runout mechanisms of landslides (Christen et al., 2010), yet the time taken for analysis and the complexities involved in modelling still makes this a challenging task (McDougall, 2017). The complexities associated with runout modelling are primarily due to the lack

CONTACT Minu Treesa Abraham phd1901204011@iiti.ac.in; minunalloor@gmail.com Department of Civil Engineering, Indian Institute of Technology Indore, Madhya Pradesh, India

of guidance in this regard. The selection of the runout model and clear guidance on modelling for practitioners are provided in very few codes or guidelines (Lato et al., 2016). The runout analysis is still considered a speciality service that demands expert support (APEGBC, 2012). Considering these facts, the practitioners need much simpler approaches for hazard assessment.

Even though there are empirical and statistical correlations used to estimate the landslide deposition area and runout distances, the relationship demands prior information on landslide volume in most cases (McDougall, 2017). Two widely accepted relationships in this regard are the inverse relationship between the reach angle and landslide volume and the one between volume and area of landslides, using Galileo scaling laws. Zhao et al. (2022) studied the empirical equations for estimating the runout distances of landslides and have argued that such may not be useful for regions other than the one from which data is collected. They have proposed a Bayesian method for estimating runout distance from sparse data, using drop height and the slope angle. In another study, a data-driven framework has been developed to predict the runout distance of landslides, using slide width, slide length, slide volume, slide thickness, and vertical drop (Xu et al., 2019). They have compared five different algorithms and have found that multi-layer perceptron performs better than the other algorithms considered. (Mergili et al., 2019) combined the release and runout in landslide susceptibility modelling using probability density functions and cumulative distribution functions of the travel distances and angles of reach of the historical landslides. The method provides approximate the susceptibility of any point in a landscape, to be affected by either shallow landslide processes or the resulting failure triggered debris flows, either through release, or through runout (Lima et al., 2019). The exportability of this model to any other study area is subject to a long process of precise data collection and statistical analysis, and it requires the thorough knowledge of multiple probabilities, to arrive at an integrated susceptibility index. Similar to this study, Melo et al. (2019) have also combined both failure and runout of shallow landslides using logistic regression and a cellular automate model.

This study is an attempt to bypass the intricacies associated with numerical modelling and existing correlations using a data-driven approach. As an initial step, a simple tool is introduced, that requires only topographical features derived from satellite-based information to estimate the maximum runout length, maximum width, maximum travel distance, and area affected due to a landslide, when the source area is identified. The method is an integration of geomorphological and geometrical landslide runout assessment techniques. The variables to be estimated were quantified for historical events using geomorphological assessment. Three different landslide typologies are considered (namely flows, slides and falls) and the data from two test sites are used for training and testing the model. The model uses a Random Forest (RF) algorithm, and the trained model is used to develop a user-friendly tool that can be easily used for applications in landslide hazard assessment. The method is tested for two different study areas in the Western Ghats of India.

2. Study area

The proposed methodology is tested at two different locations in the Western Ghats of India. The Western Ghats is a mountain range running through the Western coast of India, with a stretch of 1,600 km. The mountain range highly influences the monsoon weather patterns in India and is one of the hotspots of biodiversity in the world (Myers et al., 2000). The Western Ghats is separated into two parts by a mountain pass called the Palghat gap (Figure 1). In this study, two regions in the Western Ghats, one on the northern side and one on the southern side of the Palghat gap, are considered for the analysis. Both districts belong to the state of Kerala. The boundaries of the study areas are determined by the administrative division (district), but in the case of both the study areas, the administrative boundaries coincide with the geographical boundaries defined by the hills and valleys of Western Ghats as well. In August 2018, extremely heavy rains triggered landslides and floods in Kerala, leading to a recovery need of 4.4 billion US Dollars (United Nations Development Programme, 2018). Idukki and Wayanad were the worst hit due to landslides, and the backward socio-economic conditions of these districts also put them in a highly vulnerable condition. Owing to the higher number and catastrophic effects of landslides in these regions, quantitative hazard assessment and identification of elements exposed to risk is the need of the hour.

Idukki and Wayanad are major tourist spots in the state, and tourism and agriculture are the major income source of the inhabitants. Idukki has a relatively flatter area in the western part of the district, and the remaining locations are covered by highlands. Wayanad, on the other hand, has hills along the district boundaries. An east-flowing river Kabani and its tributaries have contributed to the landscape development of Wayanad. The tributaries of the river originate in the hilly regions on the western side and flow downhill to the lower elevation parts on the northeastern side of the district.

The lithology of both the regions is made up of rocks of the migmatite group, and the peninsular gneissic complex. Both Idukki and Wayanad contribute to the major forest cover of Kerala, with a forest area of 3151 km^2 and 1580 km^2, respectively (Forest Survey of India, 2019). Owing to the thick forest cover, both regions are rich in forest soil of high organic content.

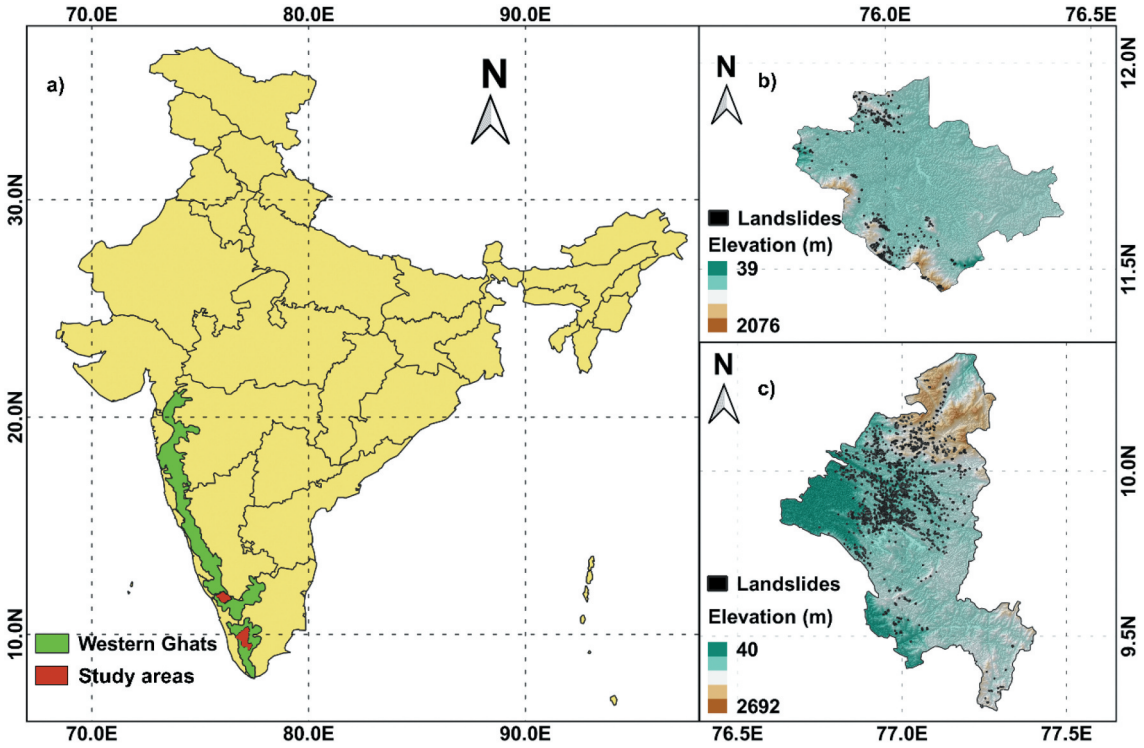

Figure 1. Location of study areas. a) India, b) Wayanad, and c) Idukki.

The midlands of these districts have lateritic soil cover, formed by the transportation of weathered rock. Riverbanks have alluvial deposits. While Wayanad and the northern parts of the Palghat gap are characterised by thick regolith deposits, Idukki has a lesser thickness of overburden soil. The high elevation ranges of both the districts are highly dissected and have witnessed deep-seated movements in history. Major debris flows extending up to a few kilometres have happened in both districts. In Idukki, the road networks in the district where the unsupported vertical slopes are highly affected by landslides. In the case of Wayanad, most landslides have happened within the forest areas.

From the interaction with the local people of the study areas, it was understood that tension cracks or new streams are usually observed prior to failure in the landslide location, particularly in the case of failure-triggered debris flows. This study is an attempt to estimate the landslide dimensions based on the topographical features of the source area using a simple and easy-to-use tool. Once the tension cracks are identified, the proposed tool can estimate the runout area, thus helping to identify the elements exposed to risk and taking necessary precautions and emergency measures.

3. Methodology

The study proposes a user-friendly tool for estimating landslide dimensions based on a data-driven approach. A detailed landslide inventory was prepared using Google earth images (Abraham et al., 2021a, 2021b) for both the study areas, and the procedure is mentioned in the 'data collection' section. The landslides were categorised into shallow landslides, flows, and rockfalls to train the models separately (Varnes, 1978). The topographical details were collected using the digital elevation models (DEMs) of the study areas (Alos Palsar DEM, with 12.5 m resolution (ASF DAAC, 2015)), and the maximum length, maximum width, area affected, and maximum horizontal distance was measured for each landslide polygon to get the training and testing data. The prepared data were used for training a model using the RF algorithm and the trained model was then used to develop a user-friendly tool for estimating landslide dimensions. The steps involved in the methodology are explained in detail in the following sections.

3.1. Data collection

Preparation of landslide inventories is the principal data required for the development of the model. For both Idukki and Wayanad, the major landslide disaster that happened in 2018 was considered for the analysis. Landslides happened throughout the higher elevation regions of both the districts in 2018, and satellite images before and after the disaster were available for preparing the inventory. The fast vegetation regrowth in the region might have resulted in missing some of the events, yet the prepared inventory was found to be in good agreement with the point landslide inventory prepared using field investigations and

satellite data interpretation (Hao et al., 2020). The process of preparing the inventory is shown in Figure 2.

A total of 2162 landslides from Idukki and 388 landslides from Wayanad were mapped using the approach mentioned in Figure 2. After locating and mapping, the typology of landslides is evaluated in detail by interpreting the google earth images. Based on the type of failure, the landslides can be classified into five (Figure 3a). The initial classification by Baltzer in 1875 (A, 1875) had only three categories: fall, flow, and slide, and this was later modified with the addition of topple and spread.

As explained in Figure 3b, the slope fails when the driving forces exceed the resisting force, that is, when the values of *FS* fall below 1. The shape of the slip surface, the material involved, and the topographical conditions decide the post-failure motion. From the collected data, no cases of topples and spreads were detected, and the inventory was classified into falls, slides, and flows. The flows are characterised by long runout, often channelised and flowing towards a stream downstream. The flow-like landslides are usually composed of both soil and rock, and the material can be classified as debris. Such flows are failure triggered, with a translational or rotational slide at the crown area and then progressing as a flow due to very high moisture content. Even though they are complex failures, including both slide and flow, the term 'flow' is

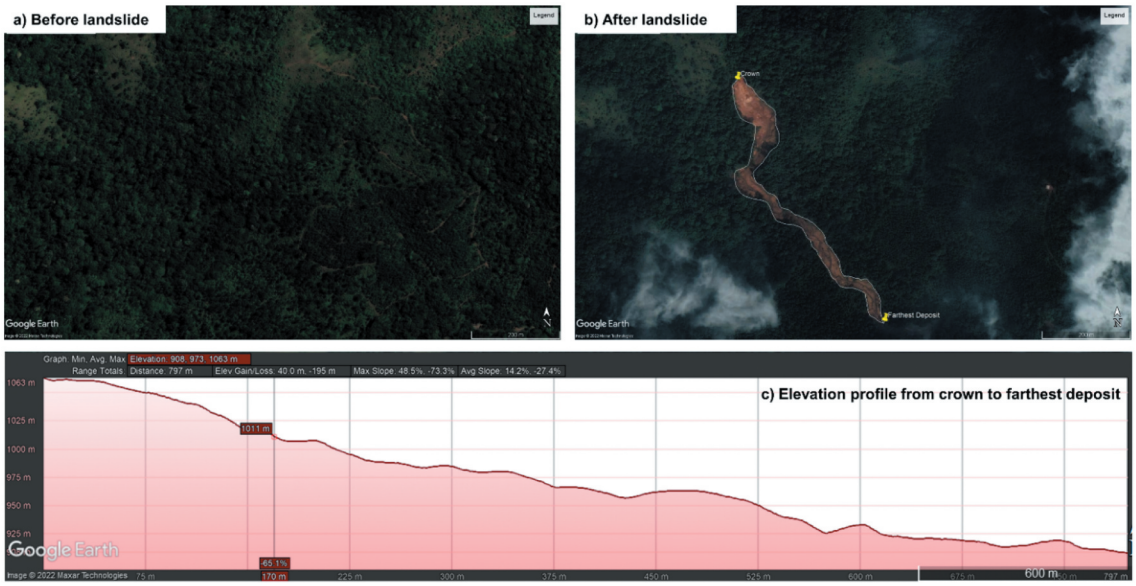

Figure 2. Preparation of landslide inventory data from pre and post landslide Google Earth Images. a) Image before the landslide, b) Image after the landslide, and c) Elevation profile along the landslide body.

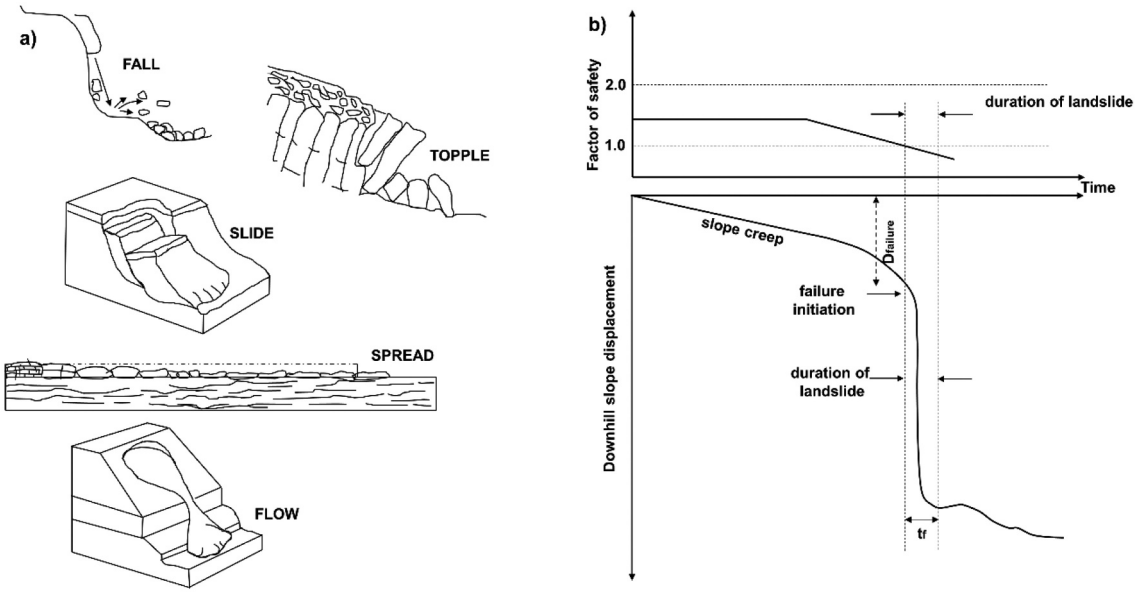

Figure 3. Types of failure and post-failure movement. a) types of failure (modified after ; Cruden & Varnes, 1996))and b) Illustration of landslide process with respect to time (modified after ; Terzaghi, 1950)).

used to classify such landslides, indicating the post-failure motion. Slides were identified where earth or debris are exposed, with lesser runout and clear distinction of a failure plane. The failure of rock in the study areas is often characterised by a complex form of both sliding and fall. The disintegrated particles of rock fall and travel longer distances and hence the combined slides and falls of rock are categorised as 'falls' in this study.

A total of 12 datasets were prepared with the collected data for training the model. This includes four sets of data for each study area, one for each landslide type and one without separating the landslide type. Apart from the separate dataset for each study area, one common dataset is also prepared for each landslide type and one superset of all the landslides from both the study areas.

These were named I1, I2, I3, IC, W1, W2, W3, WC, C1, C2,C3, and *CC*, where the first letter stands for the region and the second one stands for the type of failure considered. The letter *I* stands for Idukki, *W* stands for Wayanad, and *C* stands for the combined dataset. In the second part, 1 represents flows, 2 represents slides, 3 represents falls and *C* represents the combined dataset.

3.2. Terminology and selection of features

The main objective of the tool is to minimise the number of features used for estimating landslide dimensions. As the soil of both the study regions is of varying grain size, debris or soil is involved in both slides and flows, while the category falls includes rock particles only. Apart from the material, the topographical features play a critical role in deciding the dimensions of a landslide. The path of flow-like landslides is highly influenced by the topography, ridges, and valleys. From the visual interpretation of Google Earth images, the travel distances and the area (*A*) affected by landslides are measured manually, as shown in Figure 4. The term *L* denotes the projected runout length measured through the centre of the landslide body in the plan, and the term *W* denotes the maximum width of the cross-section of the failure. Even though *W* is also used in this manuscript to represent the region Wayanad, the name of the dataset will always be used along with a second letter or number, representing the type of failure, and hence both can be distinguished easily. The projected distance in the plan between the source and the farthest deposit location is termed the projected travel distance, and in this study, it is denoted as *D*, as shown in Figure 4. Using *D* and *W*, a rectangular bound can be proposed, which can be considered as the maximum area that can get affected by the landslide. The drop height (H) is defined as the projected distance between the source and the farthest deposit location on a vertical plane. Other features in Figure 4 are defined (by ; Hungr et al., 2005) as reach angle (α), shadow angle (β), source-talus angle (Ψ), and substrate angle (γ). The term D_x is the component of *D* in the global direction parallel to the slope. Apart from these features, the tangential angle made by the slope area to the vertical plane is denoted as θ. For different types of slopes, Finlay et al. (1999) have proposed the expressions to calculate *D* as a function of vertical drop, slope angle, volume and width of landslide. In the cases except for cut slope and boulder fall, prior knowledge of landslide volume is required to estimate the travel distance.

The extent of landslides is highly affected by the geometrical features, but the distances and angles about which the information is available before the landslide can only be considered for predicting the landslide dimensions. The features *H*, α, θ and *c* can be calculated from the DEM data and can be used for the estimation of *D*,*W*, *L*, and *A*. Both *H* and α require the knowledge of the nearest flatter area or approximate possible runout area. In the case of flow and falls, the post-failure movement will not stop at the first drop. To summarise, five input features are used, and

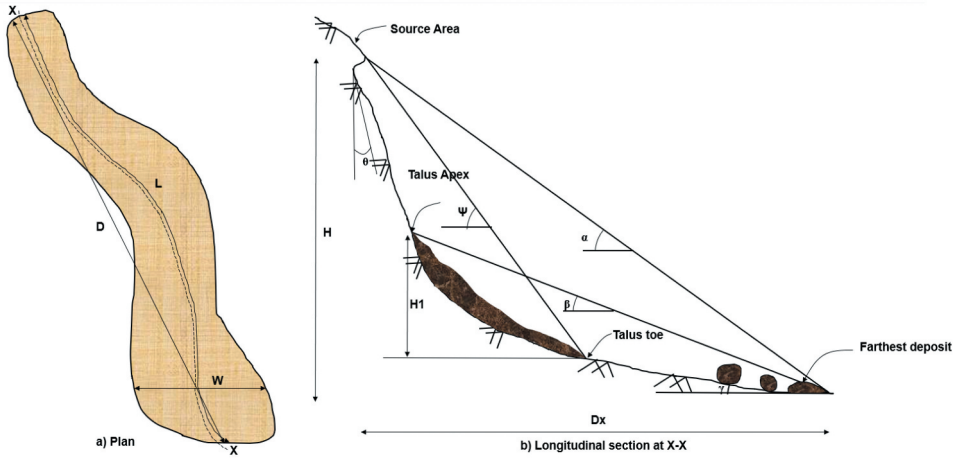

Figure 4. Geometrical features associated with landslides: a) plan, and b) longitudinal section (modified after ; Hungr et al., 2005)).

it is explored in detail if the dimensional parameters can be predicted effectively using these variables.

The distribution of all four variables in different datasets is summarised in Figure 5. It can be understood that the A, L and H values are much higher for flows in Wayanad, when compared to all other cases. In the case of W, slides have higher values than other datasets. W has the least varying distribution among all the variables.

For each of the 12 datasets prepared, three different trials were conducted, by varying the combination of input features, to predict all four variables ($A, L, W,$ and D). In the first combination, all five features are considered and is named *EHθac*. The feature importance values of each feature are used to understand the significance of features in the combination. Feature importance calculates a score for all input features, which represents the significance of each feature. The value of feature importance can vary from 0 to 1, and a higher value indicates that the specific feature has more effect in predicting the variable. As the objective is to minimise the number of features, based on the feature importance values obtained for the first combination, two more combinations were considered, one with E, H and θ, and the last one with only two features, E and θ.

Each dataset was trained and tested separately using the RF algorithm to find the best-suited features for estimating landslide dimensions.

3.3. Machine learning algorithm and performance evaluation

RF is a widely used ensemble machine learning (ML) algorithm (Ho, 1995). As the name indicates, a large number of decision trees are involved in the decision-making process of RF. Each tree in an RF has multiple branches and nodes. At each node, a decision is taken, which leads to one of the branches. The decisions thus continue, considering all the features, and the tree finally assigns a class to the object. Each tree will have a separate prediction, and later, the final prediction is decided based on voting, considering the predictions of all decision trees (Figure 6). Each decision tree is sampled independently using statistical bootstrapping (Breiman et al., 2003) and contains a subset of the dataset considered.

RF is widely used for multiple applications to train models and is proven to provide satisfactory results due to the random selection at nodes. The method is ideal for minimising the overfitting issues. The performance of the model can be further fine-tuned by varying the set of hyperparameters. The performance of the model is highly sensitive to the values of hyper parameters (Daviran et al., 2021), and several cost-effective ways are available, for multi-criteria optimisation (Liu et al., 2017). The number of trees in the forest, the maximum number of features considered for splitting a node, the maximum depth of the tree,

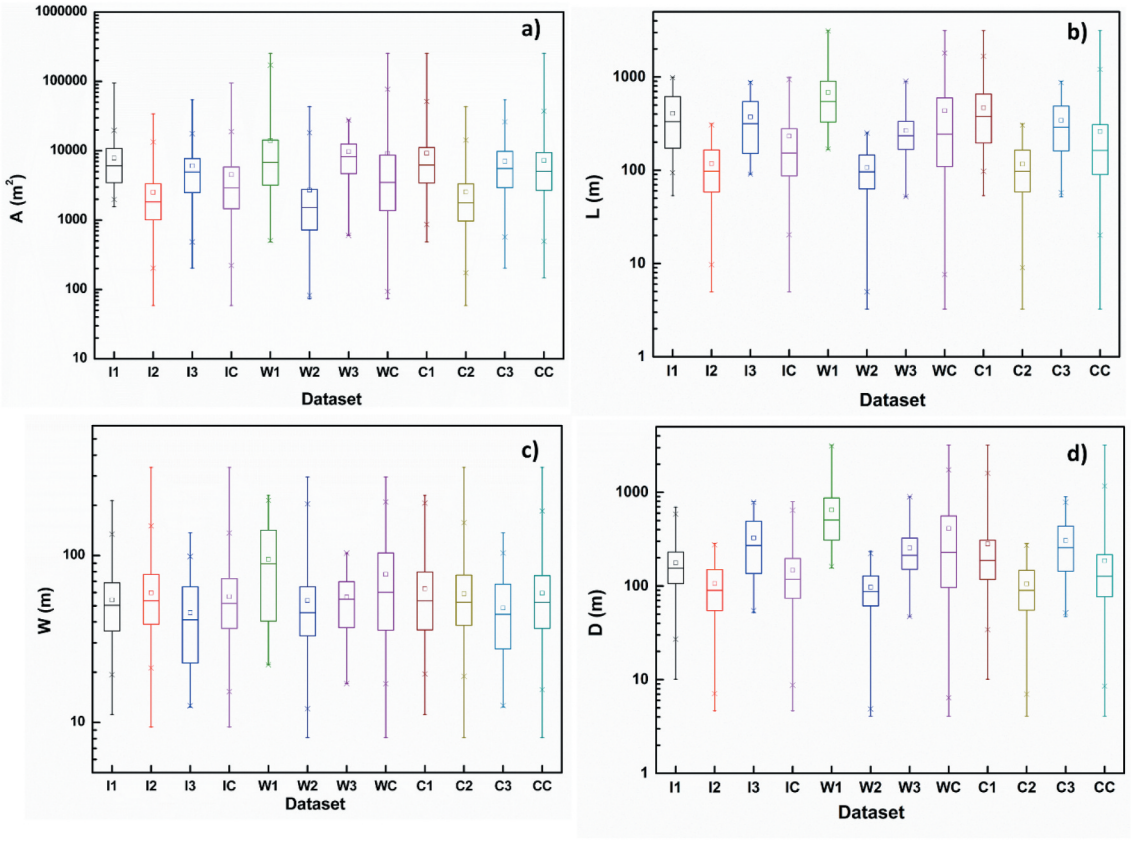

Figure 5. Box and whisker plot showing the distribution of A, L, W and D in different datasets.

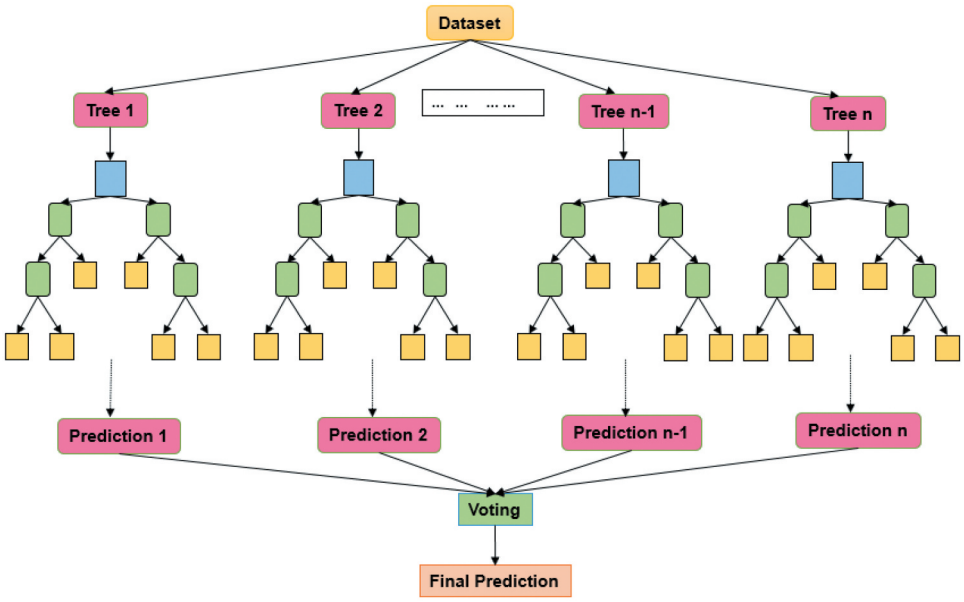

Figure 6. Graphical representation of RF algorithm.

and the minimum number of samples used to split a node are varied in this study to improve the efficiency of the model. These values were fine-tuned separately for each model, along with the test-to-train ratio of the dataset. The process is carried out manually, by varying the parametric inputs and observing the corresponding model performance. All the other parameters are constant while fine-tuning one parameter. In this study, RF regressor is used to predict the variables A, L, W and D using three different combinations of input features. The test-to-train ratio was fixed for each of these variables, based on the performance of the *EHθac* model. This is to ensure that the comparisons of errors are made on the same dataset. The ratio was varied from 0.1 to 0.5 in the case of the *EHθac* model, and the best performing test dataset was used to test the performances of the other two cases as well. Different trials were conducted to understand the effect of each feature, and the Willmott's index of agreement (*d*), Mean-Absolute-Error (*MAE*) and the Root-Mean-Square-Error (*RMSE*) values of the predicted and observed values in the test dataset were used to evaluate the performance of different models. These values are calculated based on the test and predicted data, using the following equations:

$$d = 1 - \frac{\sum_{i=1}^{n}\left(y_{\text{test,i}} - y_{\text{pred,i}}\right)^2}{\sum_{i=1}^{n}\left(\left|y_{\text{pred,i}} - \overline{y_{\text{test}}}\right| + \left|y_{\text{test,i}} - \overline{y_{\text{test}}}\right|\right)^2} \quad (1)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left(\left|y_{\text{pred},i} - y_{\text{test},i}\right|\right) \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{\text{pred},i} - y_{\text{test},i}\right)^2} \quad (3)$$

where n is the total number of samples in the test dataset, $y_{\text{test},i}$ is the variable in the test dataset and $y_{\text{pred},i}$ is the corresponding value predicted by the model. Also, $\overline{y_{\text{test}}}$ indicates the mean value in the test dataset. The value of d varies from 0 to 1, and the higher the value, the better the agreement between observed and predicted values. Based on the d, *MAE* and *RMSE* values, the features were decided and the tool for landslide dimension estimation was developed by using the best-suited feature inputs.

4. Results

Considering the regional-specific and failure-specific datasets and different feature combinations, a total of 144 models were considered for the comparative analysis. As expected from the geometrical properties, when more specific input features are available, the prediction performance will be improved. But the possible information available prior to failure is highly limited to the scarp zone, where tension cracks are limited. When multiple flat areas are available along the failure propagation, finding out the values of H and θ is challenging. The first step is to understand the importance of each of these features in predicting the variables. The least important features can be removed from further analysis. When all the features have similar importance values, it is crucial to understand the sensitivity of each feature by understanding the variations that may happen in the prediction performance if it is removed.

4.1. Feature importance and performance evaluation

The feature importance values of all 144 combinations are shown in Figure 7. The values represented in Figure 7 are the feature importance values obtained for the fine-tuned models. The maximum depth of trees, number of trees, minimum number of samples used to split a node, and the number of features to consider when looking for the best split were fine-tuned, keeping default values for all other parameters. Considering the lesser number of parameters and trials, the fine-tuning was carried out by varying the values manually, for each dataset, and each feature combination. The value of maximum depth of trees were varied from 5 to 500, with an increment of 50, and it was observed that the performance of training data went increasing, while the performance of test

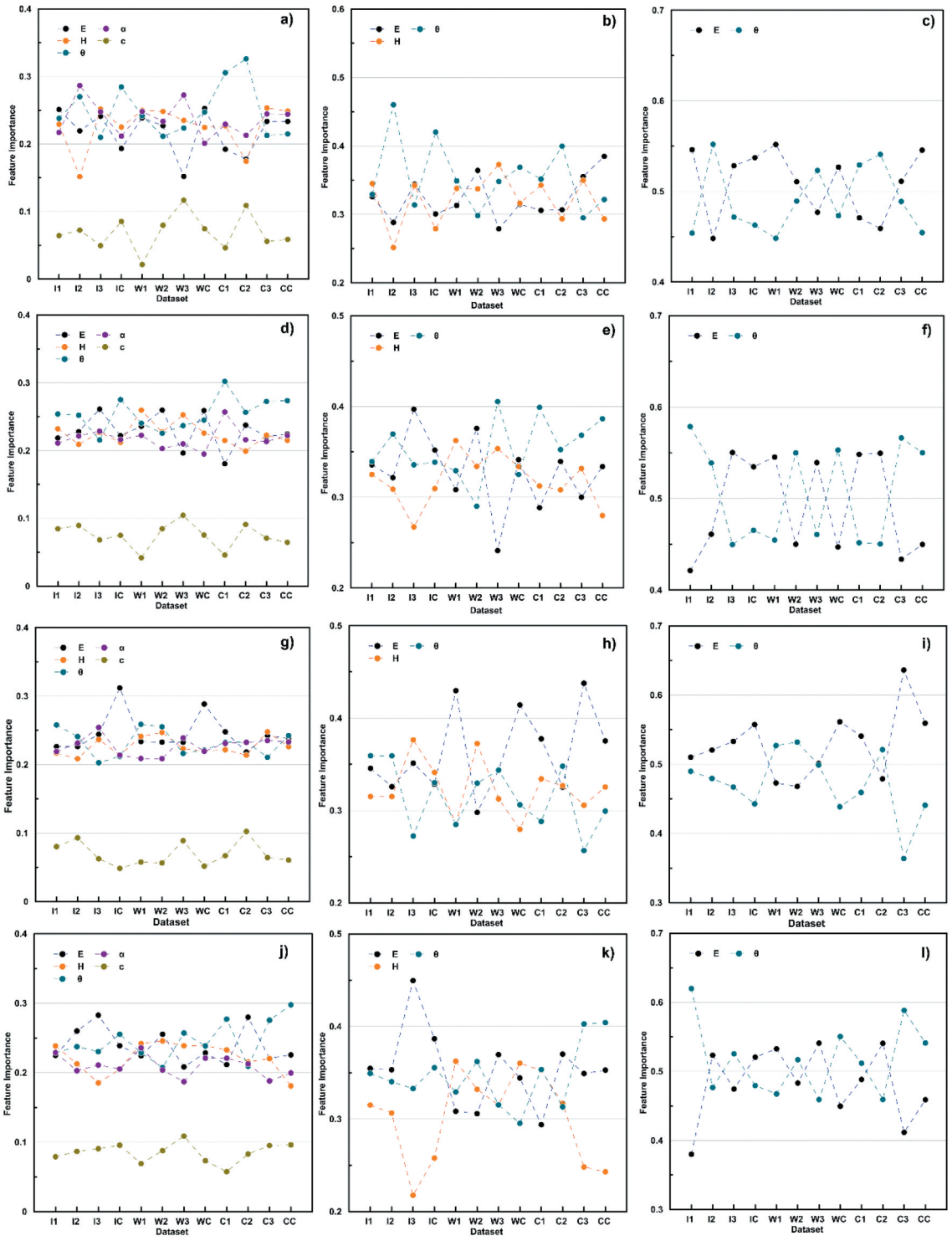

Figure 7. Feature importance values for all features for 12 datasets considered for predicting a) *A*, using *EHθac*, b) *A*, using *EHθ*, c) *A*, using *Eθ*, d) *L*, using *EHθac*, e) *L*, using *EHθ*, f) *L*, using *Eθ*, g) *W*, using *EHθac*, h) *W*, using *EHθ*, i) *W*, using *Eθ*, j) *D*, using *EHθac*, k) *D*, using *EHθ*, and l) *D*, using *Eθ*.

data was decreasing beyond the fine-tuned value. For the number of trees, the values were varied from 50 to 1000, with an increment of 50. The observation was similar for both training and testing datasets, and the fine-tuned value is the value beyond which no significant improvement in model performance is noted. The number of samples at each node were varied from 2 to 10, with an increment of 1. For the number of features, the values were varied between 1 and 3, with an increment of 0.5. All these increments were further reduced once the performance of test data becomes constant or starts decreasing. Fine intervals were used in such cases to decide the model parameters. All 144 models considered in this study have separate fine-tuned parameters, and the feature importance value, *MAE* and *RMSE* of the fine-tuned models were further used for comparison. As shown in Figure 7, the curvature values are least important in all the trials conducted using *EHθac*. While all other features have feature importance values greater than 0.15, the corresponding values for *c* were found to be less than 0.1 in most cases. The value has slightly gone above 0.1 only in the case of *W*3 and *C*2.This indicates that the feature *c* is less significant in predicting the variables, and the results may not get affected highly even if this input is avoided from the analysis. Hence, *c* is removed from further analysis.

While considering the features, the feature importance values of *E* and *θ* were found to be more than 0.3 in some cases. The effect of *θ* was found to be more significant in the case of *C*1 for predicting *A* (0.30), *L* (0.30) and *D* (0.28), *C*2 for predicting *A* (0.32), *C*3 and *C*4 for predicting *L* (0.27 and 0.27) and *D* (0.28, 0.30). Similarly, the effect of *E* was significant in the case of *IC* (0.31) and *WC* (0.29) for predicting *W*, and *I*2 (0.26), *I*3 (0.28) and *C*2 (0.28) for predicting *D*. In the case of *W*3 for predicting *A*, the feature *E* had the second least importance value of 0.15, where α was found to be the most crucial with a feature importance value of 0.27.

From the results, it was observed that *E* and *θ* have significant effects in predicting the landslide dimensions, particularly when combined datasets are used. This has special significance, as the model has to be exported to multiple regions after a detailed testing procedure, and this is the first step in developing a globally applicable model. Hence, when the model is applied to a new region for which trained models are not available. It is suggested to use the combined dataset for predicting the dimensions. This has led to the inference that *E* and *θ* cannot be avoided while developing a model. The variables *H* and *α* decide the value of *D*. If both *H* and *α* are known, *D* can be easily calculated using the geometry without using any prediction model. The challenge in this regard is the difficulty in understanding the value of *α*. When the terrain has a *D* value extending up to a few kilometres, minor variations in the expected farthest deposition point will not make much variations in α value, but this is significant in the case of slides. Estimating α prior to failure is an almost impossible task, yet most of the existing correlations use α as an input parameter for predicting *D*. From the feature importance values of the combination *EHθac*, it was observed that α is highly significant in only a few cases, and in all the other cases, it is neither the least nor the most significant factor. Hence, in the second combination, α was also removed along with c.

From the second combination, it was observed that *E* and *θ* have very high feature importance values in some cases, while *H* has values comparable to the other two, even when it is the most significant factor. Also, *H* is least important in predicting *D* in the case of *I*3, *C*3 and *CC*. Even though drop height can be estimated from the profile of the slope before failure, and it is challenging when the slope has multiple landings in between. It is difficult to use the value of *H* before failure. A wrong value can lead to large variations in predicted and observed dimensions of landslides. Hence, the third attempt was the combination of only *E* and *θ*, which can be easily calculated once tension cracks are observed. The feature importance values in this combination indicate that both the features have importance values close to 0.5 and are more or less equally contributing to the prediction of variables, in most cases except *C*3 for predicting *W*, and *I*1 for predicting *D*. In the case of *C*3 for predicting W, θ was found to be more significant with an importance factor of 0.64 and in the second case, *E* was found to be more significant with an importance factor of 0.62.

Apart from deciding the combinations, the prediction performance of all three combinations should be evaluated in detail. This helps in understanding the applicability of the model for the intended purpose. As the *EHθac* combination has more controlling parameters, it is expected to provide better prediction performance. If the results of the other two combinations are highly varying from the performance of this combination, the model cannot be used effectively with less number of features. Even though *H* and α were not the most significant factors in most cases, and their importance values were comparable with those of *E* and *θ*. Hence, removing these parameters can critically affect the predicted values. This has been evaluated using *MAE* and *RMSE* values, as shown in Figure 8. Along with the error values in the first combination (*EHθac*) represented using the line diagram, the percentage variation of the error values obtained by using the second (*EHθ*) and third (*Eθ*) are represented as a scatter plot in Figure 8.

From Figure 8, it can be observed that the errors and significantly large for *W*1 and *WC*, where long runout debris flow with very large areas are dominating. The increased values of error are in proportion with the values in the database and hence are

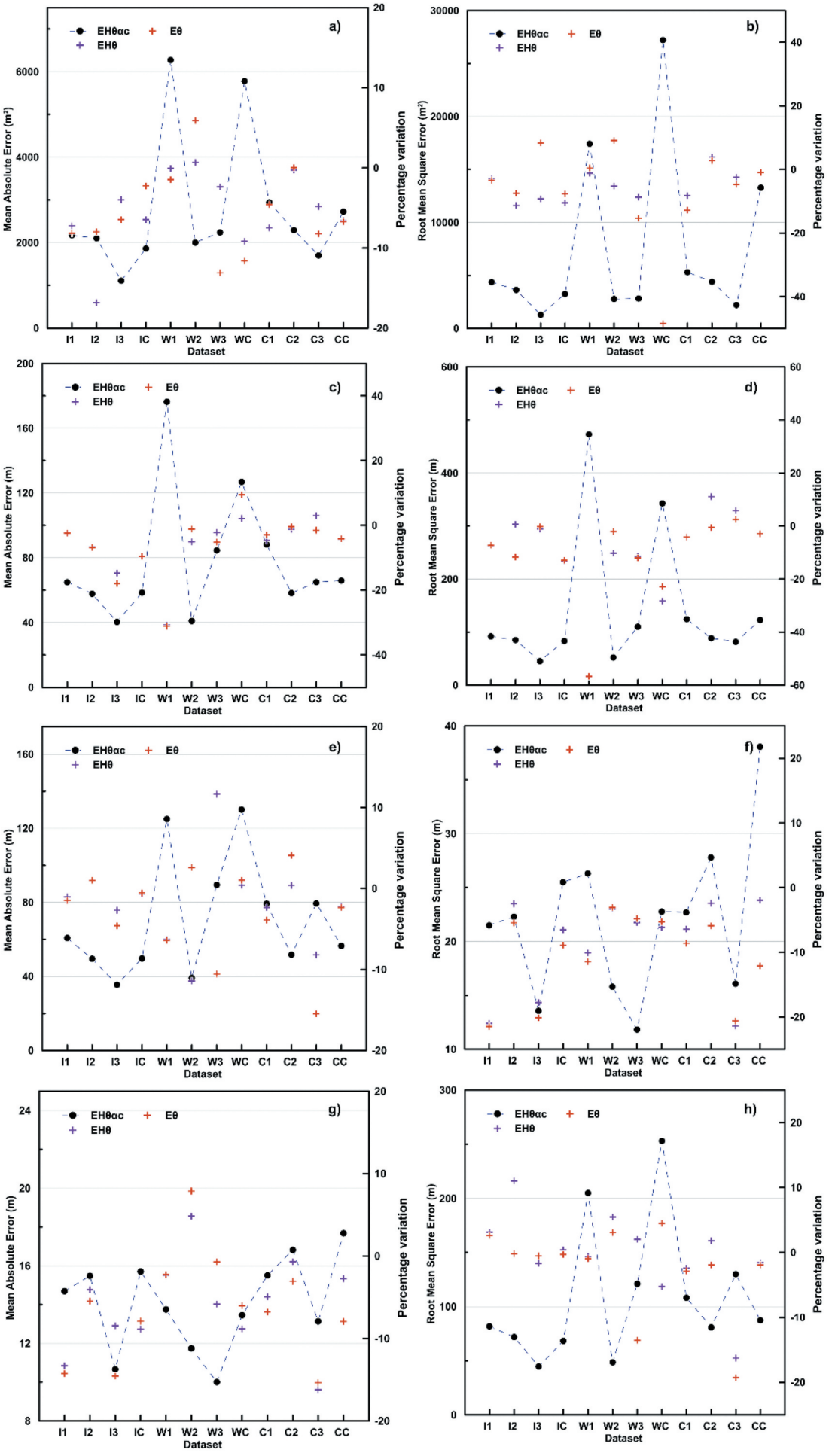

Figure 8. *MAE* and *RMSE* values in predicting different variables. a) *MAE* in predicting *A*, b) *RMSE* in predicting *A*, c) *MAE* in predicting *L*, d) *RMSE* in predicting *L*, e) *MAE* in predicting *W*, f) *RMSE* in predicting *W*, g) *MAE* in predicting *D*, and h) *RMSE* in predicting *D*.

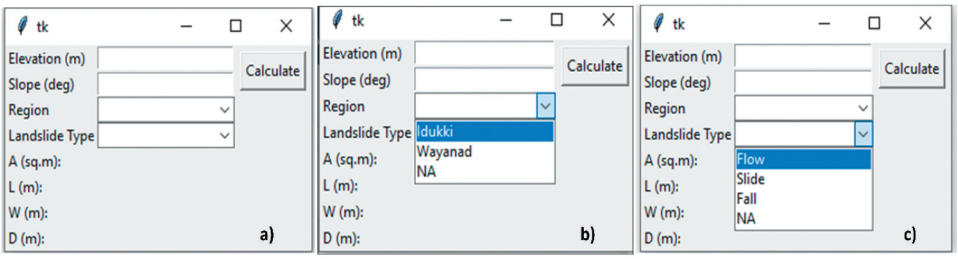

Figure 9. The interface of the proposed tool. a) All inputs, b) dropdown for region, and c) dropdown for landslide type.

acceptable. The primary concern here is the variation in the error values with the change in the input feature. When the percentage variation is negative, it indicates that the error is decreased and; therefore, the performance is improved after removing the features. When the variation is positive, it has a negative impact on the prediction. After evaluating all the cases, it can be understood that the prediction is affected by the change in features, and the maximum value of percentage variation was observed, −56.92%, in the case of estimating *L* of *W*1. The reduction has happened while using *EHθ* combination. The corresponding reduction in error using *Eθ* combination is −56.71%. The *RMSE* values of *L* in these two cases are 203.65 m and 204.64 m, respectively. This indicates that the error has been reduced considerably when the α and *c* are removed. The maximum increase in error is 11.63%, in the case of *MAE*, for predicting *D* for *W*3 dataset, using *EHθ*. The corresponding variation while using *Eθ* is −10.57%. The results indicate that in no case the error has increased beyond 11.63%, even after removing the features. This is an acceptable limit, particularly in the case of long-runout events like falls and flows. Based on the results, *Eθ* combinations of all cases were saved into pickle files as predictors to develop an interactive tool for predicting the landslide dimensions.

4.2. Description of the tool

The tool is designed with the objective of delivering an easy-to-use platform for the practitioners to estimate the area that may get affected by landslides and take necessary actions. The tool is completely developed in python environment, using the existing library functions. The regression is carried out using scikit-learn (Pedregosa et al., 2011), and the trained models are saved using pickle (Van Rossum, 2020). The tool has an interface developed using tkinter (Moore, 2018), which requires the elevation and tangential slope of the region as input features (Figure 9).

The two subsequent inputs decide which model should be used for predicting the results. The user can select Idukki and Wayanad in the present version, and if the region is outside these two, the NA option can be selected. Similarly, if the type of failure is known, the corresponding option can be selected, and NA can be selected if the type is unknown. When the material is soil or debris, and the moisture content is less, or when the terrain is relatively flat with predominantly cut slopes, slides can be expected. If the terrain has a very high moisture content and spring formations are observed nearby, flows can be expected, and falls can be expected only in the case of rocks. Based on the selection, the corresponding datasets among the 12 will be selected. The selection of the dataset has been decided after comparing the error patterns, as shown in Figure 10.

As observed from Figure 10, even though Wayanad has its own regional-specific database, the combined dataset has lesser values of error in most cases. With the use of the combined dataset, the error has increased while predicting width, in all the cases, with the maximum increase of 70.82% of *RMSE* value in the case of slides. The minimum increase is for *MAE*, for flow-like landslides, which is 7.55%. Also, in the case of slides, both *MAE*, and *RMSE* values have increased with the usage of the combined dataset. Hence, when landslide type is selected as slide, the trained model for Wayanad, *W*2 is used. The models with the Wayanad database are also used for predicting W when the region is selected as Wayanad. But in all other cases. The combined dataset is selected for predicting the dimensions of landslides in Wayanad to minimise the error. In the case of Idukki, the error is less when the regional-specific database is used. Hence, for Idukki, the *I*1, *I*2, *I*3 and *IC* datasets are used, according to the selection of user, and for Wayanad, the combined dataset is selected for flows, falls and when the type of failure is unknown. When the input for region is NA, the combined dataset is selected, corresponding to the failure type. When both region and landslide type are NA, *CC* dataset is used for prediction.

After selecting all the inputs, the 'Calculate' button can be used to get the outputs displayed on the screen. The output variables can be used to estimate the area that may get affected by the hazard. The tool is straightforward to be used and delivers the results within fractions of a second, as it uses pre-trained models for prediction. The performance of the tool is currently evaluated for the datasets retrieved in this

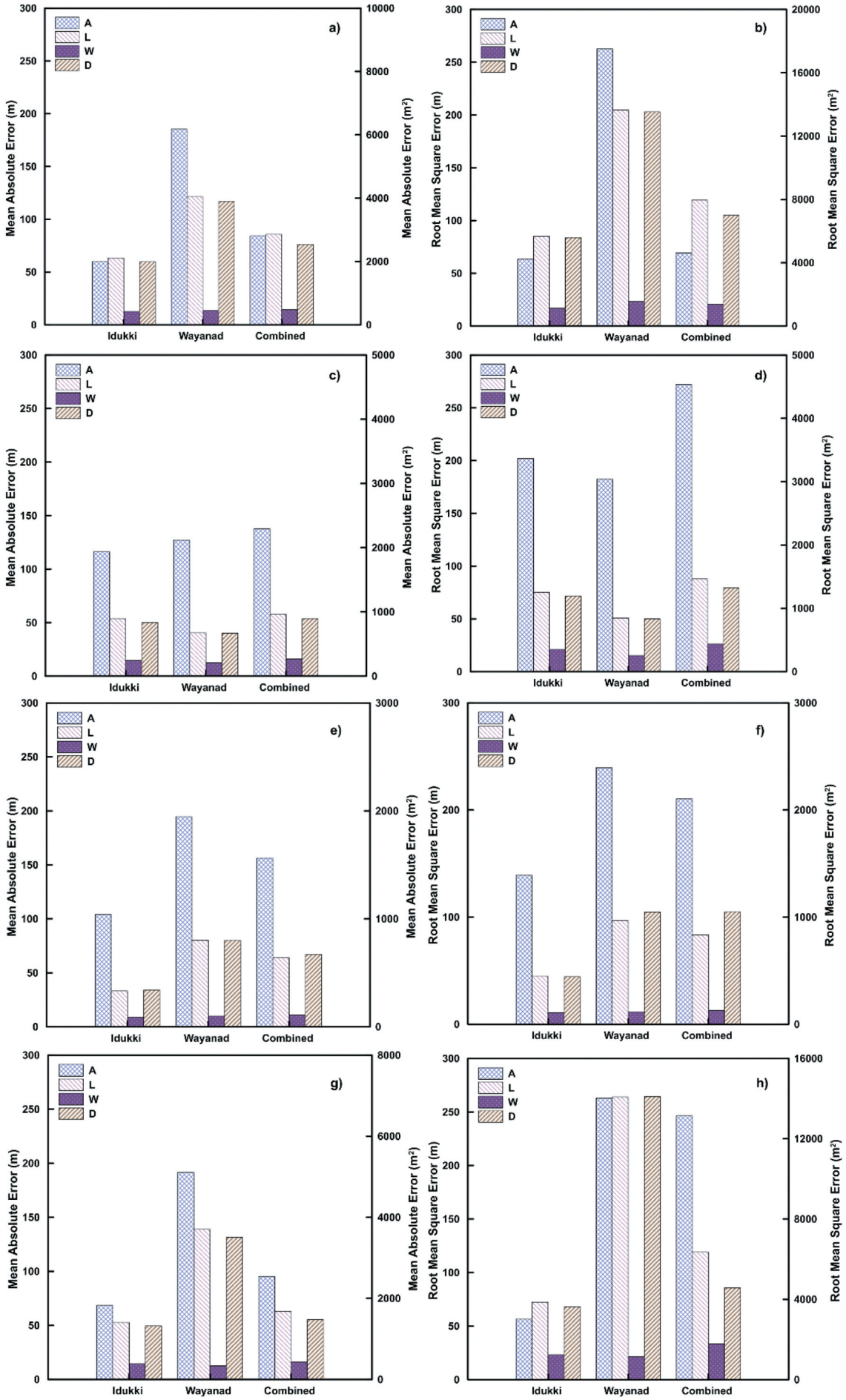

Figure 10. Comparison of *MAE* and *RMSE* values of *Eθ* combination, for different landslide types. a) *MAE* for flow, b) *RMSE* for flow, c) *MAE* for slide, d) *RMSE* for slide, e) *MAE* for fall, f) *RMSE* for fall, g) *MAE* for combined landslide types, and h) *RMSE* for combined landslide types.

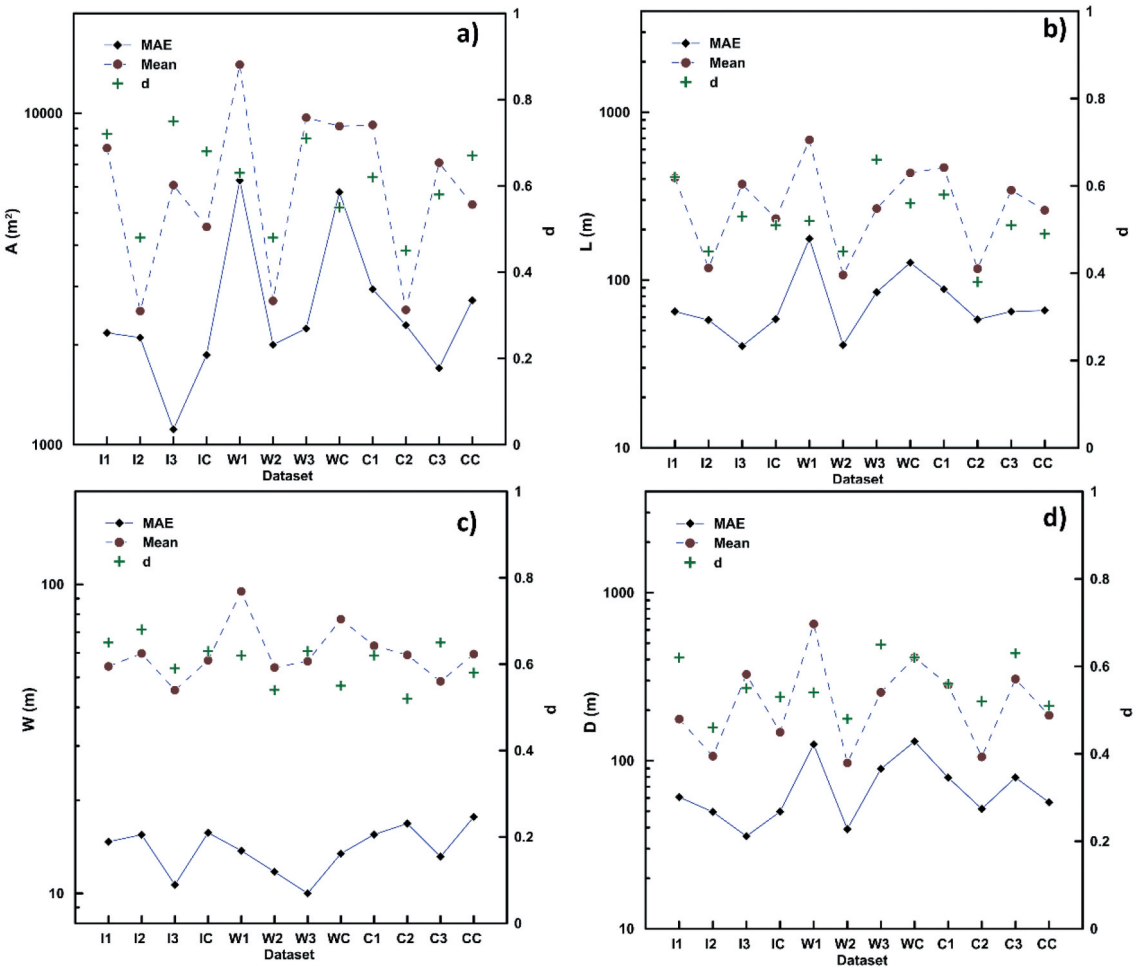

Figure 11. *MAE*, mean and *d* values of A, L, W and D in different datasets.

study only. For extending its applicability other regions, thorough analysis with regional specific data is required. While the existing literature presents statistical correlations with drop height and other parameters which are obtained after the occurrence of landslide. Such relationships have limited applicability on decision support and early warning applications. This study puts forward a set of promising results which indicate that the dimensions of landslides can be predicted using the elevation and slope information of the location of visible cracks. The method is much simpler when compared with the numerical modelling-based tools and can be used to identify the elements exposed to risk and take necessary actions before the occurrence of landslide. The primary reason for considering the tool an 'easy-to-use' one is the usage of minimum parametric inputs, with satisfactory outputs, as demonstrated by the *MAE* and *RMSE* values. While using a numerical model requires detailed knowledge of the triggering mechanism, topography, boundary conditions and material properties, this tool can be used easily with only elevation and slope information. Also, the interface provides an easy way to input the parameters and directly getting the outputs, rather than using regression equations or codes. The method can be further enhanced by collecting data from multiple regions in order to develop a globally applicable version of the tool.

5. Discussion

Study presents a data-driven approach to predict the dimensions of different landslide types using an easy-to-use tool. The results indicate that the geometrical parameters such as E, H, θ, α and c can be used for estimating the area, maximum length, maximum width and maximum travel distance of landslides. The study investigates in detail the possibilities of predicting the dimensions using the elevation and slope data alone. Figure 7 indicates that the use of the $E\theta$ combination reduces the error when compared to the $EH\theta ac$ combination, in most of the cases. While evaluating the feature importance values, it can be understood that the curvature is the least important factor in all the 144 combinations considered. This is because most landslides have concave profile curvatures, and the values are highly similar irrespective of their size. Hence, any change in curvature values will not affect the model performance. Along with curvature, the angle of reach, α, was also removed in the second analysis. Even though there was a significant reduction in error in some of the cases

using the second combination, the *MAE* value has increased by more than 11% in the case of predicting *D* with *EHθ*. Eventhough an influential parameter was removed, the performance was not highly affected, as the hyperparameters were fine-tuned separately for each model. The reason for *E* being a critical factor in estimating the dimensions is closely related to the landslides that have happened in the study area. In both the study areas, *E* values are closely related to *H*. The landslides in higher elevation zones have happened in forest areas and plantations, where the *H* values are also higher. The failures in lower elevation regions are induced by the cut slopes exposed without lateral support. Their dimensions are controlled by the construction activities such as buildings and roads, and they usually have lesser drop heights. Thus, even though *H* values are removed from the input features, *E* values indirectly represent *H*. The reason for using *E* instead of *H* is that *E* value can be collected when the source area is known, while *H* value cannot be. This limits the applicability of the model to different regions, where *E* and *H* are not closely related. Thus, the methodology has to be tested before exporting to other regions. The number of trees, features at each node, minimum number of samples used to split a node and depth of trees were used to minimise the error in each case. A similar approach was adopted and was found satisfactory in reducing the error for *Eθ* combination as well. This helped in developing a model by using only slope and elevation data for predicting landslide dimensions.

While comparing the errors, it should be noted that the $AL, W,$ and *D* values are entirely different for each landslide type. From Figure 11, it can be observed that the *MAE* values in all cases are lesser than the mean values of the dataset. While flows and rockfalls have very long runouts extending from a few hundreds of metres to a few kilometres, slides extend up to a few hundreds of metres only, with an average value close to 100 m (Figure 11). The *MAE* values in the case of the area are more than 2000 m^2 in all the cases, except falls for Idukki and the combined dataset. The *MAE* values in predicting the length of the slides are also very close to the mean values of the dataset. This has happened due to the considerable variation in the size of slides. The error values in prediction will overestimate the landslide hazard in the case of small landslides. This limitation can be bypassed by preparing a different dataset for cut slope failures and shallow landslides in other locations. While shallow slides happening in forest regions have a wider travel distance, the failed mass in case of cut slope failures often gets deposited at the foot of the slope due to the flat area nearby. In such cases, the failure plane is primarily vertical, while in the case of other shallow landslides, circular or translational slip surfaces inclined to the horizontal plane are observed. Considering *d*, the values are least in case of slides, and the minimum value observed is 0.38, in the case of slides in the combined dataset. In all cases except slides, the values are greater than 0.5, showing satisfactory agreement between the observed and predicted datasets.

The width of landslides is maximum for both slides and falls, and the values range from 8 m to 337 m. The error values should be evaluated with reference to the distribution of data in each case. From Figure 8 it can be understood that in the case of area, the maximum *MAE* is with *W*1, which is 6269.79 m^2. This error has highly influenced the combined dataset of Wayanad as well, which has an *MAE* value of 5778.57 m^2. Even though the magnitude is higher when compared to the other values, the predictions are satisfactory, as the area of debris flows in this region ranges from 485 m^2 to 253,880 m^2. The pattern of variation of error is similar for $A, L,$ and *D*, with the maximum error in the case of *W*1 and *WC* datasets. But in the case of width, the maximum error is observed in the case of *I*2 and *IC* datasets. This is due to the higher number of cut slope failures that happened in the Idukki district. The failures are very wide and have a width values upto 337 m, with lesser values of length. Also, there are shallow landslides that happened away from the road, which have width-to-length ratio close to unity. The error is slightest in the case of rockfalls, where the average width is 48 m in the case of a combined fall dataset, and the values vary from 12 m to 137 m. These observations indicate that when the dataset is uniform, there are higher chances that the error will be minimum. The hypothesis during formulating the methodology was that when the same type of failure happens in a nearby location, the dimensions will not vary much, and hence a model trained using historical data can effectively be used for predicting future events in the same area.

In the case of the Wayanad dataset, this hypothesis is not valid, as the error is lesser for the combined dataset than the regional-specific Wayanad dataset. The main reason for this variation is among the 388 landslides mapped from Wayanad, 252 are flows, 68 are slides, including both earth and debris, and the remaining are falls. The number of events that is being used for training is much less when compared to the whole dataset, and the variation within the dataset is also very high. Due to these reasons, the model gets better trained with combined datasets. Hence, the regional-specific datasets need not be the best option while predicting landslide dimensions, and the number and quality of data plays critical roles in minimising the error. With more data, the model can be extended to other parts of the world as well.

The methodology can only be used to predict the dimensions of a landslide that may happen in the

future, based on the field observations of cracks. However, no information can be provided on the time of occurrence of the landslide. The tool should be the regional or local scale landslide early warning systems to obtain the information on 'when' a landslide will occur. The warning system can be based on rainfall thresholds, seismic signals, or satellite or field-based monitoring systems. The integration of this tool along with an early warning system can provide a better understanding of the hazard and can be used to disseminate the warnings effectively to the stakeholders. This aspect can be explored in the future. In the present state, the proposed methodology is best suited for long runout failures like flows and falls and can be used to develop a globally applicable model for predicting landslide dimensions.

6. Conclusions

The study presents a data-driven approach to predict the dimensions of landslide, upon the identification of minor cracks in the crown area. The methodology proposed in this study use only the elevation and tangential slope of the crown area to predict the dimensions of landslides and prove to be a promising tool that can be used in the decision support system.

The proposed methodology is tested for two different regions in Western Ghats of India, using 12 different datasets, and three different combinations of input features are used to evaluate the influence of each parameter on the model predictions. The comparison of *MAE* and *RMSE* values of the predicted variables in each case indicates that the maximum increase in error is only 11.63%, while the reduction in error is 56.92% with the $E\theta$ combination. The performance of $E\theta$ combination was found to be comparable with the other two, without the requirement of any challenging feature inputs like drop height and angle of reach.

The combinations of all 12 datasets were used to decide the model to be used for predicting the dimensions. While the region-specific models were found to have least errors for Idukki, the combined dataset was found to have better performance than the datasets for Wayanad. The pre-trained models were used to develop a tool with an interactive interface, which can be easily used to predict the landslide dimensions. The proposed methodology has the potential to be applied to other regions as well with the availability of regional specific data, yet the possibility of finding robust relationships among the variables should be evaluated through detailed analysis.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data availability statement

The data used for the analysis is available on request from the corresponding author.

Tool availability

The tool mentioned in the manuscript will be shared on request by the corresponding author.

References

A, B. (1875). *Über bergstürze in den Alpen*. Verlag der Schabelitz'schen buchhandlung (C. Schmidt). Zurich.

Abraham, M. T., Satyam, N., Jain, P., Pradhan, B., & Alamri, A. (2021a). Effect of spatial resolution and data splitting on landslide susceptibility mapping using different machine learning algorithms. *Geomatics, Natural Hazards and Risk*, *12* (1), 3381–3408. https://doi.org/10.1080/19475705.2021.2011791

Abraham, M. T., Satyam, N., Lokesh, R., Pradhan, B., & Alamri, A. (2021b). Factors affecting landslide susceptibility mapping: assessing the influence of different machine learning approaches, sampling strategies and data splitting. *Land*, *10*(9), 989. https://doi.org/10.3390/land10090989

APEGBC. (2012). *Professional practice guidelines – Legislated flood assessments in a changing climate in British Columbia* (Vancouver, British Columbia: Engineers and Geoscientists British Columbia). 2012.

Armento, M. C., Genevois, R., & Tecca, P. R. (2008). Comparison of numerical models of two debris flows in the Cortina d' Ampezzo area, Dolomites, Italy. *Landslides*, *5* (1), 143–150. https://doi.org/10.1007/s10346-007-0111-2

ASF DAAC, 2015. *Alaska Satellite Facility Distributed Active Archive Center (ASF DAAC) Dataset: ASF DAAC 2015*, ALOS PALSAR_Radiometric_Terrain_Corrected_high_res; Includes Material © JAXA/METI 2007. [WWW Document]. https://doi.org/10.5067/Z97HFCNKR6VA

Breiman, L., Last, M., & Rice, J. (2003). Random Forests: Finding Quasars Feigelson, E.D., Babu, G. J. *Statistical Challenges in Astronomy* (pp. 243–254). Springer-Verlag. https://doi.org/10.1007/0-387-21529-8_16

Christen, M., Kowalski, J., & Bartelt, P. (2010). RAMMS: Numerical simulation of dense snow avalanches in three-dimensional terrain. *Cold Regions Science and Technology*, *63*(1–2), 1–14. https://doi.org/10.1016/j.coldregions.2010.04.005

Cruden, D., & Varnes, D. (1996). Landslide types and processes. In A. K. Turner & R. L. Schuster Eds., *Landslides, investigation and mitigation* (pp. 36–75). Transportation Research Board. Special Report.

Daviran, M., Maghsoudi, A., Ghezelbash, R., & Pradhan, B. (2021). A new strategy for spatial predictive mapping of mineral prospectivity: Automated hyperparameter tuning of random forest approach. *Computers & Geosciences*, *148*, 104688. https://doi.org/10.1016/j.cageo.2021.104688

Finlay, P. J., Mostyn, G. R., & Fell, R. (1999). Landslide risk assessment: Prediction of travel distance. *Canadian Geotechnical Journal*, *36*(3), 556–562. https://doi.org/10.1139/t99-012

Forest Survey of India. (2019). State of forest report 2019. In *Dehradun* 2 . Uttarakhand: 131–140 .

Gariano, S. L., & Guzzetti, F. (2016). Landslides in a changing climate. *Earth-Science Reviews*, *162*, 227–252. https://doi.org/10.1016/j.earscirev.2016.08.011

Guzzetti, F., Crosta, G., Detti, R., & Agliardi, F. (2002). STONE: A computer program for the three-dimensional simulation of rock-falls. *Computers & Geosciences*, *28*(9), 1079–1093. https://doi.org/10.1016/S0098-3004(02)00025-0

Hao, L., van Westen, C., Martha, T. R., Jaiswal, P., & McAdoo, B. G. (2020). Constructing a complete landslide inventory dataset for the 2018 monsoon disaster in Kerala, India, for land use change analysis. *Earth System Science Data*, *12*(4), 2899–2918. https://doi.org/10.5194/essd-12-2899-2020

Ho, T. K., 1995. Random decision forests, in: *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*. pp. 278–282. https://doi.org/10.1109/ICDAR.1995.598994

Huang, H., Huang, J., Liu, D., & He, Z. (2021). Understanding the public responses to landslide countermeasures in southwest China. *International Journal of Disaster Risk Reduction*, *64*, 102500. https://doi.org/10.1016/j.ijdrr.2021.102500

Hungr, O., Corominas, J., & Eberhardt, E. (2005). Estimating landslide motion mechanism, travel distance and velocity. In O. Hungr, R. Fell, R. Couture, & E. Eberhardt (Eds.), *Landslide risk management* (pp. 30). Taylor and Francis Group, CRC Press.

Lato, M., Bobrowsky, P., Roberts, N., Bean, S., Powell, S., Stead, D., McDougall, S., Brideau, M. A., & VanDine, D., 2016. . In Canadian technical guidelines and best practices related to landslides: a national initiative for loss reduction 8114 (Geological Survey of Canada)https://doi.org/10.4095/299117 .

Leroueil, S., Locat, J., Vaunat, J., Picarelli, L., Lee, H., & Faure, R. (1996). Geotechnical characterization of slope movements. In K. Senneset (Ed.), *7th International Symposium on Landslides* (CRC Press) (pp. 53–74).

Lima, P., Steger, S., Netto, A. L. C., Glade, T., & Mergili, M., 2019. Combining landslide susceptibility with potential runout. An integrative approach combining data-driven methods., in: *IAG Regional Conference on Geomorphology 2019*. 19-21 September (International Association of Geomorphologists) Athens, p. 536.

Liu, C. H. B., Chamberlain, B. P., Little, D. A., & Cardoso, Â., 2017. Generalising random forest parameter optimisation to include stability and cost, in: Y. Altun, K. Das, T. Mielikäinen, D. Malerba, J. Stefanowski, J. Read, M. Žitnik, M. Ceci, & S. Džeroski (Eds.), *Machine Learning and Knowledge Discovery in Databases; Proceedings of European Conference, ECML PKDD 2017 Skopje, Macedonia, September 18–22, 2017 Proceedings, Part III, Lecture Notes in Computer Science*. Springer International Publishing, Cham, pp. 102–113. https://doi.org/10.1007/978-3-319-71273-4_9

McDougall, S. (2017). Landslide runout analysis — Current practice and challenges. *Canadian Geotechnical Journal*, *54*, 605–620. https://doi.org/10.1139/cgj-2016-0104

Melo, R., Zêzere, J. L., Rocha, J., & Oliveira, S. C. (2019). Combining data-driven models to assess susceptibility of shallow slides failure and run-out. *Landslides*, *16*(11), 2259–2276. https://doi.org/10.1007/s10346-019-01235-2

Mergili, M., Schwarz, L., & Kociu, A. (2019). Combining release and runout in statistical landslide susceptibility modeling. *Landslides*, *16*(11), 2151–2165. https://doi.org/10.1007/s10346-019-01222-7

Moore, A. D. (2018). *Python GUI Programming with Tkinter: Develop responsive and powerful GUI applications with Tkinter*. Packt Publishing.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, *403*(6772), 853–858. https://doi.org/10.1038/35002501

Pedregosa, F., Gaël Varoquaux, A. G., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Terzaghi, K. (1950). Mechanism of landslides Paige, Sidney. In *Application of geology to engineering practice*. Vol. (Geological society of America), pp.83–123https://doi.org/10.1130/Berkey.1950.83 doi:

United Nations Development Programme. (2018). *Kerala post disaster needs assessment floods and landslides-august 2018* (United Nations Development Programme) 1–440.

Van Rossum, G. (2020 (Python Software Foundation.)https://www.python.org/downloads/release/python-382/). The python library reference, release 3.8.2.

Varnes, D. (1978). *Slope movement types and processes*. Transp. Res. Board Spec. Rep.

Xu, Q., Li, H., He, Y., Liu, F., & Peng, D. (2019). Comparison of data-driven models of loess landslide runout distance estimation. *Bulletin of Engineering Geology and the Environment*, *78*(2), 1281–1294. https://doi.org/10.1007/s10064-017-1176-3

Zhao, T., Lei, J., & Xu, L. (2022). An efficient Bayesian method for estimating runout distance of region-specific landslides using sparse data. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, *16*(1), 140–153. https://doi.org/10.1080/17499518.2021.1952613