Supplementary material

Table of Contents

Supplementary Material A

Steps to idealise an ellipse from a landslide polygon

Here we outline the six steps for calculating the length-to-width ratio of an idealised ellipse from a landslide polygon. These provide further detail and justification to supplement **Tables III and IV** (main text).

Step A: Preliminary landslide spatial attributes. For elliptic approximation Methods 1 to 3 (**Table III**, main text), we first use existing GIS tools to calculate the centre of gravity and orientation of the original landslide shape, which has a landslide area (A_L) and perimeter (P_L). The centre of gravity is based on a weighted mean of the landslide shape's Northing and Easting coordinates (N, E) (Khorshidi, 2009) and forms the map coordinates that define the centre point of the final best-fit elliptic shape. In a small number of cases, it is possible for the centre of gravity to fall outside of the original landslide shape's extent. The orientation of the original landslide is calculated in a GIS, which is used to align the landslide best-fit ellipse's long axis.

Step B: Fit convex hull. For the elliptic approximation Method 2 and Method 3, a convex hull (CH) is fit to the original landslide shape using dedicated tools within a GIS. A convex hull is the smallest shape that completely contains the perimeter of the original landslide shape, where all internal angles of the CH are less than 180° (de Berg *et al.*, 2008). The area (A_{CH}) and perimeter (P_{CH}) of the CH are then used as the inputs to the following Step C. Within the GIS, A_{CH} is calculated using the shoelace algorithm (ESRI, 2016) and P_{CH} using Pythagorean theorem (Prashker, 1999).

Step C: Calculate ellipse length-to-width ratio from area and perimeter of the original landslide shape or the CH fit. In order to 'simplify' the asymmetric and irregular shape of landslide polygons to a regular ellipse, we can use the area and perimeter of the original landslide shape (A_L, P_L) or the CH fit to the landslide shape (A_{CH}, P_{CH}), to estimate the ellipse length-to-width ratio ($\Lambda_E = L_E/W_E$) with the same area and perimeter as the original or CH shape.

The length-to-width ratio Λ_E for an idealised elliptic shape is calculated using the solution to a quadratic equation, based on the equations for ellipse area and perimeter. The area of an ellipse A_E, is calculated as (Heath, 1931),

$$A_E = \pi ab \quad \text{Eq. (S1)}$$

where a = semi-major axis of ellipse, b = semi-minor axis of ellipse. Calculating the exact perimeter of an ellipse (P_E) is complex, so it is common to use a variety of elliptic approximations based on the long (a) and short (b) axes of the ellipse. Here, due to its relative simplicity, we use Euler's (1773) formula (described in Michon, 2015),

$$P_E \approx \pi\sqrt{2(a^2 + b^2)} \quad \text{Eq. (S2)}$$

where P_E = ellipse perimeter, a = semi-major axis of ellipse, b = semi-minor axis of ellipse. We note that this approximation for P_E can be up to 10% inaccurate for very long, thin ellipses (Michon, 2015). By combining **Eqs. S1** and **S2**, we can solve for the length-to-width ratio of an ellipse (a/b) = (L_E/W_E), in terms of the area (A_E) and perimeter (P_E) of the ellipse, by using the solution to a quadratic equation, giving,

$$\Lambda_E = \frac{L_E}{W_E} = \frac{4\pi A_E}{P_E^2 - \sqrt{P_E^4 - 16\pi^2 A_E^2}} \quad \text{Eq. (S3)}$$

We then substitute into **Eq. S3** the area (A_E) and perimeter (P_E) of the ellipse with the area and perimeter of original landslide shape (A_L, P_L), or the CH fit to the landslide shape (A_{CH}, P_{CH}), thus approximating the length-to-width ratio (Λ_E) of an idealised elliptic shape with the same area and perimeter as that of the original landslide or the CH fit.

When substituting the original landslide P_L for the ellipse P_E, **Eq. S3** becomes sensitive to the original landslide's perimeter 'sinuosity' relative to its area (A_L). If the original landslide has a 'sinuous' perimeter (P_L), this has the potential to force the idealised elliptic shape to longer, thinner forms. For this reason, we fit a CH to the original landslide shape which removes most internal voids from the landslide shape.

Step D: Calculate ellipse length and width from length-to-width-ratio and area. The ellipse length (L_E) and width (W_E) can then be expressed in terms of the length-to-width ratio of the ellipse Λ_E and the landslide area (A_L) or area of the CH fit (A_{CH}) which we denote as A:

$$L_E = 2\sqrt{\frac{A\Lambda_E}{\pi}}$$
$$W_E = 2\sqrt{\frac{A}{\pi\Lambda_E}} \quad \text{Eq. (S4)}$$

67 *Step E: Scale ellipse area to match original landslide area*. For elliptic approximation Method
68 3 the length (L_E) and width (W_E) of the ellipse approximated from a CH fit to the original
69 landslide shape is scaled so that the area of the CH (A_{CH}) matches the area of the original
70 landslide,

$$L_{Scaled\ CH} = L_{CH}\sqrt{\frac{A_L}{A_{CH}}}$$
$$W_{Scaled\ CH} = W_{CH}\sqrt{\frac{A_L}{A_{CH}}} \qquad \text{Eq. (S5)}$$

71 This scaling does not affect the 'shape' of the ellipse in terms of Λ_E, but can result in a better
72 match in terms of area between the idealised elliptic shape (A_E) and the original landslide
73 shape.

74 *Step F: Plot ellipse from length, width, centroid and orientation*. The final stage of elliptic
75 approximation Methods 1 to 3 is to plot the ellipse using standard GIS tools. The inputs to
76 the GIS tools are (i) ellipse length (L_E) and width (W_E); (ii) ellipse centroid coordinates, which
77 correspond with where the length and width segments intersect; and (iii) ellipse orientation,
78 which corresponds to the angle of orientation of the ellipse long axis L_E. The output is an
79 elliptical approximation of the original landslide shape, which should approximately align with
80 and overlay the original landslide.

81

82

Supplementary Material B:

Kolmogorov-Smirnov testing of distribution of landslide ellipse length-to-width ratio (Λ_E) between landslide area categories.

In this supplementary material we describe the use of a Kolmogorov−Smirnov (K−S) two-sample test (Lillefors, 1967) to compare the distribution of Λ_E for each of the two landslide inventories in the ten landslide area categories, numbered 1 to 10 with 1 being the smallest landslide area category and 10 the largest (**Figure S1**, left to right along the *x*-axis), using the following hypothesis:

H_0 The Λ_E values in landslide area category *i* come from the same statistical one-point probability distribution as the Λ_E values in landslide area category *j*, where *i* and *j* represent the ten different categories of landslide areas given in **Figure S1**, $i \neq j$, $i = 1,2,\ldots,10$; $j = 1,2,\ldots,10$.

H_1 The Λ_E values in landslide area category *i* do not come from the same statistical one-point probability distribution as the Λ_E values in landslide area category *j*, where *i* and *j* represent the ten different categories of landslide areas given in **Figure S1**, $i \neq j$, $i = 1,2,\ldots,10$; $j = 1,2,\ldots,10$.

In its standard form, the two-sample K−S test measures the maximum distance (*D*) between two empirical cumulative frequency curves (Lillefors, 1967). Here, we will compare the cumulative frequencies of Λ_E for each pair of A_L categories. If the maximum distance (*D*) between the two empirical curves (each based on a different landslide area category) is small, this suggests both datasets ($F_i(x)$ and $F_j(x)$) follow approximately the same statistical one-point probability distribution. If *D* is large, this suggests significant deviation between the two cumulative distribution functions (Wilcox, 2005). The maximum distance *D* is then compared to already established tables to calculate the probability, *p*, that this magnitude of *D* would have been observed if the two datasets did truly come from the same distribution. **Figure S1** shows the K−S test *p* value results of comparing each landslide area category (A_L) to each other for Northridge and then Guatemala.

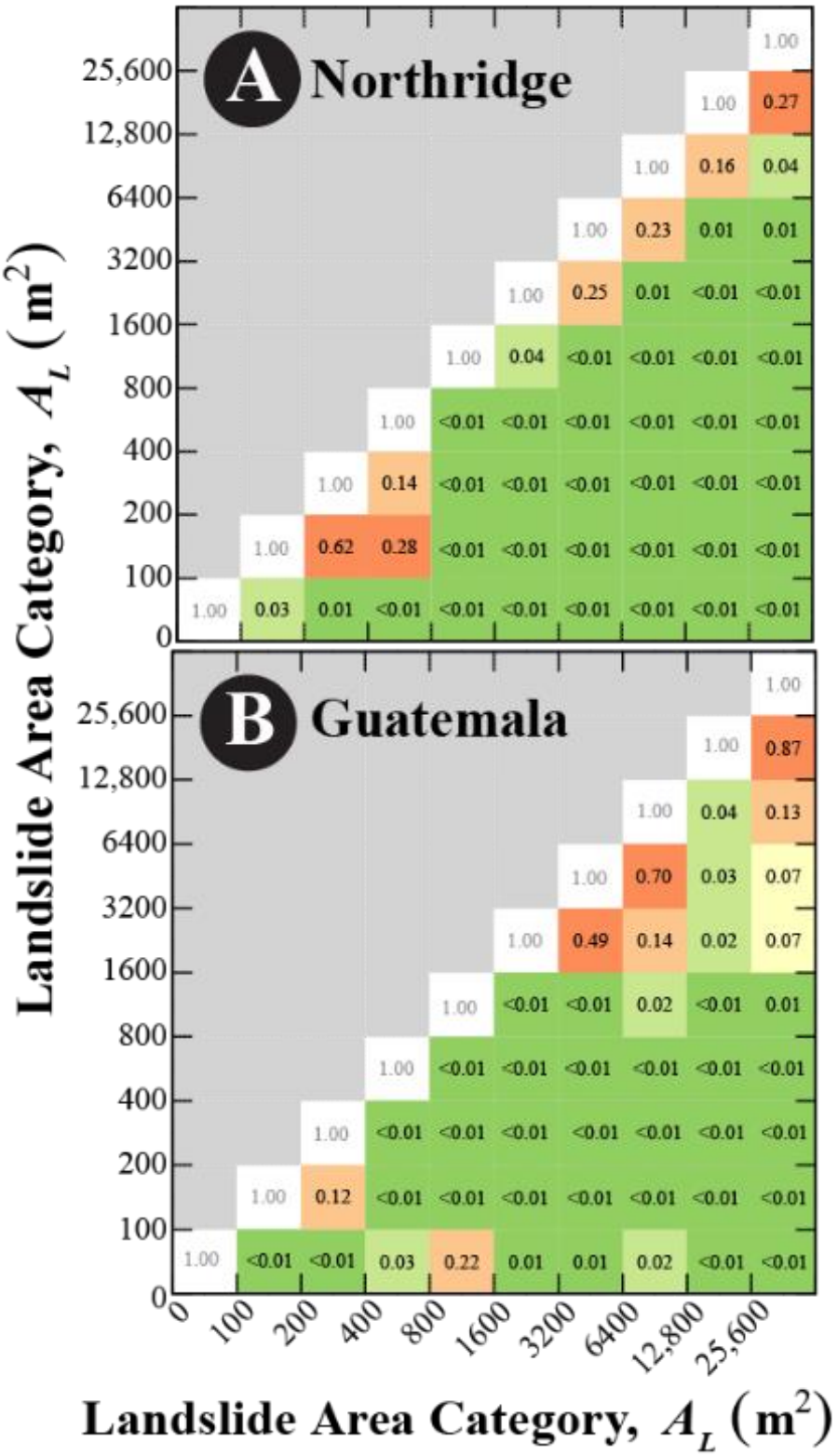

Figure S1. Probability *p* values from the two-sample Kolmogorov−Smirnov tests to compare distributions of landslide length-to-width ratio (Λ_E) between each landslide area category. **(A)** 9,441 landslides in the 1994 Northridge landslide inventory and **(B)** 8,031 landslides in the 1998 Guatemala landslide inventory. Box colours correspond to the significance of the *p*-value, where green denotes very low *p*-values (i.e., there is a low probability the samples were drawn from the same distribution) and orange denotes high *p*-values (i.e., there is a high probability the samples were drawn from the same distribution). Values of $p = 0.00$ denote very small values of *p* (e.g., $p < 10^{-3}$) rounded to 0.00.

Using a significance level of $p = 0.10$, results in 38 out 45 pairs of A_L categories where the

null hypothesis is rejected that the probability density distribution of Λ_E is the same in both A_L categories. The *p* values are generally higher in neighbouring landslide area categories for the 'tails' of the categories (i.e., very small and very large landslide areas), which is attributed to small sample sizes in these categories. Broadly there is a difference in the distribution of Λ_E between pairs of landslide area categories, with some discrepancies.

Supplementary Material C:

Modified Kolmogorov-Smirnov goodness-of-fit test for probability density functions fit to landslide ellipse length-to-width ratio within each landslide area category

In this supplementary material we describe the use of a modified version of the Kolmogorov−Smirnov (K−S) goodness-of-fit test (Lillefors, 1967). We use this to confront the empirical cumulative distributions for Λ_E (i.e., all original landslide Λ_E in each landslide area category) with the model pdf best-fit MLE parameters (i.e., the solid lines in **Figure 10** (main text) for each landslide area category).

The standard goodness-of-fit K−S test is not appropriate if the model parameters describing the distribution have been estimated from the dataset, as the significance levels stated in standard tables will be more conservative than reality (Lillefors, 1967; Crutcher, 1975; Deluca and Corral, 2013). Moreover, the K−S test is particularly sensitive to the region of the distribution where the majority of the data points lie (i.e., the mode), and thus less sensitive to the values obtained for the goodness-of-fit of the tails of the distribution (where fewer data points lie) (Mason and Schuenemeyer, 1983). Instead, the K−S test can be adapted in a Monte-Carlo type workflow to simulate what typical deviations are observed between a known distribution and random values selected from that distribution. The process outlined by Clauset *et al.* (2009) is adopted here:

1. *Fit the inverse-Gamma probability density function* (pdf) *to the observed* (empirical) Λ_E *data* for each landslide area category using MLE. This gives us parameter values ρ_{OBS}, a_{OBS}, s_{OBS} describing the best-fit pdf to the data.
2. ***Perform a 'standard' K−S test*** to calculate the *D* value, defined as the largest vertical distance (in linear coordinates) between the empirical cumulative distribution function and the model inverse-Gamma cumulative distribution function fit to the observed data. This is D_{OBS}.
3. *Randomly generate* n_{OBS} *values from the inverse-Gamma pdf* with the parameters ρ_{OBS}, a_{OBS}, s_{OBS}. The variable n_{OBS} = the number of values in the observed dataset.
4. Using MLE, fit the inverse-Gamma distribution to these randomly generated values, this gives a pdf with parameters ρ_{SIM}, a_{SIM}, s_{SIM}.
5. *Again, perform a K−S test* to calculate the *D* value, but this time between the cumulative distribution for the randomly generated values (Step 3) and the model inverse-Gamma cumulative distribution fit to these values (Step 4). This is D_{SIM}.
6. *Repeat* ***Steps 3−5*** for n_{iter} iterations where n_{iter} is related to the desired level of precision

(here, we chose 2,500 iterations). In each case, the values ρ_{SIM}, a_{SIM}, s_{SIM} will vary slightly. For each iteration, record D_{SIM}.

7. *Compare D_{OBS} to the range of D_{SIM}.* The significance level (p) is defined by the percentile of D_{SIM}, i.e., to test the data to $p = 0.90$ significance level, compare D_{OBS} to the 90^{th} percentile of D_{SIM}. If $D_{OBS} < (p \times D_{SIM})$, we can accept the null hypothesis that the observed Λ_E data come from the inverse-Gamma pdf with parameters ρ_{OBS}, a_{OBS}, s_{OBS}.

This method works on the premise that the observed (empirical) data can be considered to have come from the same distribution as that fit to it if the D value lies within the normal range of those encountered when fitting a distribution to random deviates from a known distribution. **Figure S2** shows boxplots of the distribution of D_{SIM} values and D_{OBS} for each landslide area category for both Northridge and Guatemala.

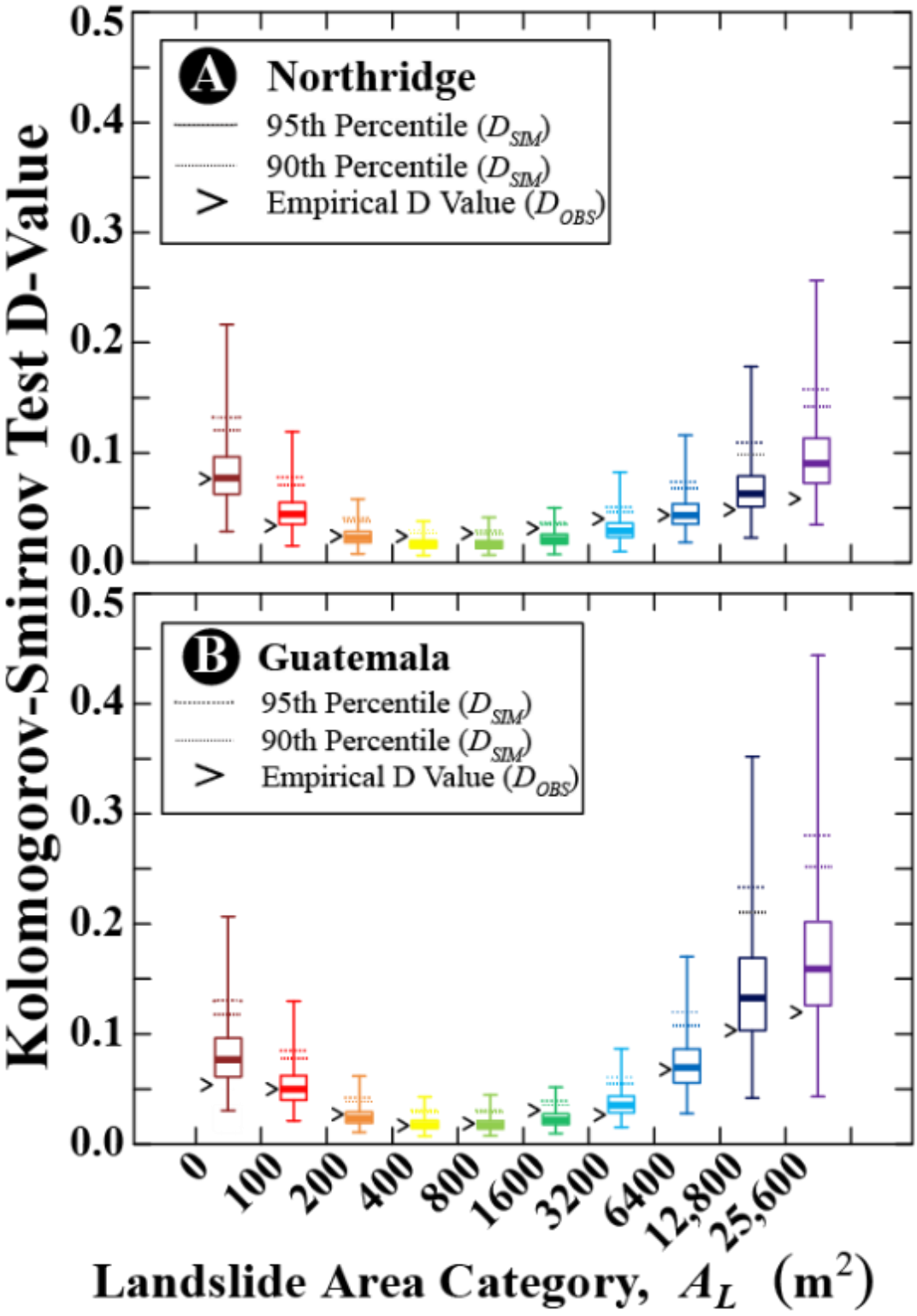

Figure S2. Boxplots of D_{SIM} values from Monte-Carlo Kolmogorov−Smirnov goodness-of-fit testing, compared to the observed D_{OBS} value (indicated by >). The upper and lower horizontal lines of the boxplot rectangles represent the interquartile range of the D_{SIM} values (25th and 75th percentiles), the thick horizontal lines the median D_{SIM} (50th percentile), the boxplot upper and lower horizontal lines the maximum and minimum D_{SIM} values. Also shown for each boxplot are the 95th percentile for the D_{SIM} values and represented by dotted horizontal lines and the 90th percentile by dashed horizontal lines.

For all landslide area categories for both the Northridge and Guatemala landslide inventories

in **Figure S2**, D_{OBS} (indicated by the symbol > next to each boxplot) is below the 95th and 90th percentile of D_{SIM} values (indicated by the dotted and dashed horizontal dashed lines, respectively) and corresponding to $p = 0.05$ and $p = 0.10$, respectively. This suggests that the deviations between observations and the best-fit pdf are within the 'normal range' to be expected if the data were drawn from that distribution. We therefore conclude from **Figure S2** that the 20 inverse-Gamma pdfs shown in **Figure 10** (main text) are reasonable models for the distribution of landslide ellipse Λ_E for both Northridge and Guatemala.

Supplementary Material D

Python Grass-GIS code for fitting an ellipse to landslide polygon

This Python code utilises Grass-GIS tools to fit an ellipse to each landslide polygon in an inventory. It then measures the goodness-of-fit of each ellipse using the ellipticity index (e_L) (**Eq. 4**, main text) and the length-to-width ratio of the ellipse (Λ_E) (**Eqs. 1-3**, main text). The workflow is shown in **Figure 4** (main text)

Available to download from:

https://github.com/faithatron/Landslide_shape_tools/

Supplementary Material E

R Code for Inverse Gamma probability density function fitting of length-to-width ratio

This R code will perform maximum likelihood estimation fitting of an inverse-Gamma probability density function (pdf) (**Eq. 5**) to landslide length-to-width ratios (Λ_E) generated using the Grass-GIS model code in **Supplementary Material D**. A bootstrapping technique (**Supplementary Material C**) is used to test the goodness-of-fit of the model pdf to the original data and add error bars to the plots of the pdf. The R code will generate a figure similar to **Figure 10** (main text) showing the best fit pdf to the observed data for each landslide area category. Comments are included in the code to explain the steps.

Available to download from:

https://github.com/faithatron/Landslide_shape_tools/