Supporting Information for "Multiscale Spatial Patterns in Giant Dike Swarms Identified through Objective Feature Extraction"

A. Kubo Hutchison[1], L. Karlstrom[1], T. Mittal[2]

[1]Department of Earth Sciences, 1272 University of Oregon, Eugene, Oregon 97403, USA

[2]Department of Geosciences, 503 Deike Building, University Park, PA 16802

Contents of this file

1. Text S1 to S3
2. Figures S1 to S4

Additional Supporting Information (Files uploaded separately)

1. Captions for Datasets S1 to S4

Introduction

Text S1.

1. Parallel Line Segments

For two parallel line segments with slope, m, and intercepts b_1 and b_2 the distance, d_c between the two lines compared to the same difference in Hough space $dHT = |\rho_2 - \rho_1|$ regardless of origin. In the following, we show this mathematically.

In cartesian space the distance between the two parallel lines is

$$d = (b_1 - b_2)sin(\theta). \quad (1)$$

where

$$\theta = arctan(\frac{1}{m}) = arctan(\frac{x_a - x_b}{y_a - y_b}) \quad (2)$$

where the subscripts a,b reflect the end points of the each respective segment.

In Hough space the distance between the two points is $|\rho_2 - \rho_1|$ where

$$\rho_1 = (x_1 - x_0)cos\theta + (y_1 - y_0)sin\theta. \quad (3)$$

$$\rho_2 = (x_2 - x_0)cos\theta + (y_2 - y_0)sin\theta. \quad (4)$$

with an origin of (x_0, y_0). The intercept of the segments, which is the only difference between them can be written as:

$$b = y_a - m * x_a \quad (5)$$

for b_1 and b_2 and each segment's respective endpoints. However, for b' modified for changing the origin the equation is

$$b' = (y_a - y_0) - m * (x_a - x_0) \quad (6)$$

Rewriting Equation 5 using rearranging Equations 2-4 we find:

$$y_a = \frac{\rho_1}{sin\theta} + \frac{x_a - x_0}{tan\theta} + y_0 \quad (7)$$

Substituting Equation 6 and 2 into Equation 5 yields

$$b_1 = \frac{\rho_1}{sin\theta} + \frac{x_0}{tan\theta} + y_0 \quad (8)$$

$$b_2 = \frac{\rho_2}{sin\theta} + \frac{x_0}{tan\theta} + y_0 \quad (9)$$

but

$$b' = \frac{\rho}{sin\theta} \tag{10}$$

Finally substituting Equations 7 and 8 into 1 we prove that

$$d_C = \rho_2 - \rho_1 = d_{HT} \tag{11}$$

the distance between two parallel lines in Cartesian space is equal to distance between the two points in the Hough space. For truly parallel line segments, there is no ρ distortion between Hough and Cartesian space and it is not dependent on angle or origin.

2. Subparallel Line Segments

For two subparallel line segments, in the same quadrant, with some small difference in angle θ_e, what is the ρ distortion and is it dependent on angle or origin choice? We analyze this case by taking two line segments (red and blue respectively in Fig S1) which are some small θ_e apart in angle with $\rho_2 > \rho_1$.

With two different angles we find the perpendiculars for each respective line.

$$\rho_2 = (x_1 - x_0)cos\theta + (y_1 - y_0)sin\theta. \tag{12}$$

$$\rho_1 = (x_2 - x_0)cos(\theta + \theta_e) + (y_2 - y_0)sin(\theta + \theta_e) \tag{13}$$

In subparallel lines, there are two distances of interest as defined by the point formed by the intersection of the perpendicular and line formed by the line segment. We define two distances d_1 and d_2 which are analogous to d above and the ratio of interest $\frac{d_1}{d_2}$.

Using the right triangles shown in the inset of Figure S1, we can find that

$$d_1 = \frac{\rho_2}{cos(\theta_e)} - \rho_1 \tag{14}$$

and

$$d_2 = \rho_2 - \frac{\rho_1}{cos(\theta_e)} \tag{15}$$

which yields the ratio

$$\frac{d_1}{d_2} = \frac{\rho_2 - \rho_1 cos\theta_e}{\rho_2 cos\theta_e - \rho_1} \tag{16}$$

which as $\theta_e \to 0$, $\frac{d_1}{d_2} \to 1$ and the distance is equal to $\rho_2 - \rho_1$.

Plugging in 12 and 13 into 14 and 15 respectively and expanding the trigonometry identity using $(\theta + \theta_e)$, we find that the origin choice does affect the d_1 and d_2 values but disappears as $\theta_e \to 0$.

$$d_1 = (x_2 - x_0)cos\theta + (y_2 - y_0)sin\theta - cos\theta_e\Big((x_1 - x_0)[cos\theta cos\theta_e - sin\theta sin\theta_e] + (y_1 - y_0)[sin\theta cos\theta_e + cos\theta sin\theta_e]\Big) \tag{17}$$

$$d_2 = (x_2 - x_1)cos\theta cos\theta_e + (y_2 - y_1)sin\theta cos\theta_e + cos\theta_e((x_1 - x_0)sin\theta + (y_1 - y_0)cos\theta) \tag{18}$$

which both simplify to $(x_2 - x_1)cos\theta + (y_2 - y_1)sin\theta$ when $\theta_e \to 0$ and the origin terms cancel out.

The distance in Hough space d_{HT} is singular and equal to :

$$d_{HT} = \sqrt{(\theta_e)^2 + (\rho_2 - \rho_1)^2} \tag{19}$$

where

$$\rho_2 - \rho_1 = (x_2 - x_0)cos\theta + (y_2 - y_0)sin\theta - ((x_1 - x_0)[cos\theta cos\theta_e - sin\theta sin\theta_e] + (y_1 - y_0)[sin\theta cos\theta_e + cos\theta sin\theta_e]) \tag{20}$$

This shows that for almost subparallel line segments, there is a distortion in ρ space of the Hough Transform which is dependent on both angle and origin location. As the angle difference betwene the two lines decreases, the dependence on angle and origin decreases.The ratio of $\frac{d_1}{d_2}$ can measure the amount of distortion in between two segments. When $\frac{d_1}{d_2}$ is near one the distortion is small and the difference in angle is small.

For the limit when θ_e is small, we can simplify Equation 20 to

$$d_{HT} = \sqrt{d_p + [(x_1 - x_0)sin\theta - (y_1 - y_0)cos\theta] * \theta_e} \tag{21}$$

3. Crossing Negative to Positive Angles

In the HT space, lines with $-90°$ and $90°$ have the same horizontal orientation (E-W from a map point of view). Thus we would wish to cluster some dikes which have this orientation. However, there is additional distortion which occurs when we cluster from negative to positive angles. We can illustrate this with an example of two points in HT space $p_1 = [-90°, 100]$ and $p_1 = [90°, 100]$. However, in Cartesian space, when we calculate the intercept between these two lines using Equation 10 the intercepts are far apart $b_1 = -100$ and $b_2 = 100$ leading to an overly wide cluster violating our desired constraints for high aspect clusters. In this case, a third point $p_3 = [90°, -100]$ would have the same orientation as p_1 since $b_3 = -100$.

To solve this issue in HT space clustering we simply rotate the dataset so that the median angle is centered on 20 degrees which aims to minimize the amount of clusters which would need to cross $-90°$ and $90°$ and $0°$.

Text S2.

4. Clustering Sensitivity Analysis

To investigate dike cluster lengths and act as data reduction we apply a clustering algorithm to the datasets transformed into Hough Space. We chose to use Agglomerative

Hierarchical clustering (Everitt, 1980) which has two parameters which effect the algorithm: first, the linkage criterion of which we chose complete linkage; and secondly the distance parameter over which clusters will not be merged. In the two dimensional space of the Hough Transform there are two different scales over which this distance parameter calculates Euclidean distance, the angle which varies from -90 to 90 and ρ value which varies over several orders of magnitude from positive to negative values and is measured in kilometers. To account for this, we scale our datasets by the θ and ρ cutoff values. For all datasets, we use the same θ threshold of 2° while the ρ threshold varies for each dataset. For the ρ threshold, we have chosen the smallest scale length available in the data which is the mean dike segment length. Our choices for these parameters are explained in the main text.

To examine the effect of choosing different parameters and the robustness of our clustering analysis we performed a sensitivity analysis on the Chief Joseph Dike Swarm dataset testing the effects of changing three clustering parameters: ρ threshold, θ threshold, and linkage criterion (Figure S2). These tests were performed with the Scipy Hierarchical clustering algorithms (Virtanen et al., 2020). When changing the ρ threshold (Figure S2 a,d,g), we applied a θ threshold of 2° using complete linkage. When changing the θ threshold (Figure S2 b,e,h), we applied a ρ threshold of 400 m using complete linkage. When changing the linkage (Figure S2 c,f,i), we applied a ρ threshold of 400 m and a θ threshold of 2°. For reference, the ρ threshold of 400 m is what was used in the main text. Each result was then put through the same post-processing steps and compared. We applied the same filtering step described in the Methods section of the main text.

We see that increasing it significantly to 1000 m has little affect on median cluster sizes although it slightly increases the outlier clusters. Increasing it again to 5000 m, a 10X, does introduce significantly larger clusters and raises median cluster significantly for the filtered datasets. As would be expected it does raise the cluster width however the ranges ranges seen in the 400 m and 1000 m cut off are similar.

We tested three linkage types; first complete linkage in which the distance is determined by the furthest most data points in a cluster; second, single linkage in distance is determined by the closest objects in a cluster only; finally, average or unweighted pair-group method aproach (UPGMA) in which the distance between two clusters is the average distance between all objects in those clusters. Single linkage is the most effected by single data points while complete linkage creates more compact clusters. There are a few extremely large clusters using single linkage ($<$ 600 segments) this is likely due to the "chaining" phenomenon in which single segments cluster and this creates more "outlier" clusters however the ranges for dike cluster length and width remain consistent. This behavior can be desirable for some research applications but not for the issue of linking segments with like orientation. After filtering for cluster size, the single linkage also reveals the largest deviations from complete and average linkage likely due to the presence of chaining clusters which are also more likely to be $>$ 3 segments. Average linkage behaved similarly to complete linkage but showed slightly larger, longer, and wider clusters.

Changing the θ threshold has little effect on median cluster size until it reaches 10° which does show significantly larger clusters in the filtered dataset and higher extremes in size. Changing angle from 1° to 2° effects the filtered clusters but has little effect on the overall dataset. Increasing it again from 2° to 4° increases the median size from 2 to

3. When it comes to cluster length, changing the θ threshold does not effect the range of cluster lengths seen in the cluster data base although it does effect the range seen in the filtered database likely due to the changing median size. This behavior would be expected since the Hough transform has no information about how segments are arranged relative to one another and changes are likely due to increasing cluster sizes. It does however increase cluster width linearly. This is described above in Text S1 since increasing the θ threshold increases the values of θ_e and increases possible widths between subparallel segments. Overall, changing θ threshold has relatively larger effects on the cluster length then changing ρ threshold.

Text S3.

5. Identification of Radial Swarms

Due to the distortion described above and the nature of the Hough Transform as a set of dikes gets further from the choosen origin of the Hough Transform it will appear more and more like a radial swarm despite being a random array of orientations. This is a function of the size of the radial swarm, it's range and standard deviation of ρ rather than it's Cartesian length. To test this we created several completely random dike swarms with no radial pattern of different scales (1, 10, 100 km) and placed them at different distances from the origin then fit the radial swarm equation (Eq. 9 in the main text) to the datasets. To evaluate how well it is fit, erroneously, by a radial swarm we look at the goodness of fit (R_{sq}) value (Figure S3). We find that when the distance from the HT origin is scaled by the standard deviation in ρ (μ), radial swarm start to appear erroneously at approximately $2 - 2.5\mu$. Assuming the dike swarms are normally distributed in ρ this indicates that the majority of dikes ($> 95\%$) which will fall at distances less than $\pm 2\mu$ can be correctly identified as radial if that is there pattern and few swarms will be incorrectly identified as radial. Additionally, using this knowledge one can move the HT center accordingly to account for features that are in the far ranges of ρ if necessary and run the radial identification as many times as necessary then compile it into one Cartesian dataset.

Data Set S1. Linked dike clusters for the Columbia River Flood Basalt group including the four identified subswarms: Chief Joseph, Monument, Ice Harbor, and Steens as compiled in Morriss, Karlstrom, Nasholds, and Wolff (2020). This dataset was produced using the Agglomerative Clustering algorithms using the parameters set in Table 1. This dataset is in the format of a CSV file but can be read into GIS programs using Well Known Text (WKT) linestring. This dataset uses the a UTM Zone 11N projection (EPSG:26911). The file includes the start and end points of the average line in the cluster and it's mid points, cluster length and width (Xstart, Xend, Xmid, Ymid, in meters and UTM coordinates, Dike Cluster Width or R_Width, Dike Cluster Length or R_Length all in meters); calculated average ρ and θ for the Hough Transform ρ units measured in meters, θ units measured in degrees, unless otherwise stated); the origin used for the Hough Transform which is different for each subswarm (xc,yc, meters in UTM coordinates); average slope and intercept (AvgSlope, AvgIntercept meters); range and standard deviation for ρ and θ for all objects in the cluster (ρ units measured in meters, θ units measured in degrees); cluster size (Size); sum of segment lengths in a cluster (SegmentLSum, meters); whether the cluster crosses between negative and positive values (ClusterCrossesZero, boolean); overlap as calculated in the main text where the length of overlap is normalized by the sum of segment lengths in a cluster; maximum number of overlapping segments (nOverlapingSegments); twist angle which is the difference in angle betweeen the average cluster line and the average line formed by cluster midpoints (EnEchelonAngleDiff, degrees); the p-value for the midpoint line fit of the segments where $p < 0.05$ is considered to be a significant fit (EEPValue); the maximum, median, and minimum segment nearest neighbors distances in the cluster which is calculated using the cartesian midpoints of each segment and normalized by the Cluster Length (MaxSegNNDist, MedianSegNNDist, MinSegNNDist); characterization of each cluster as filtered or not, filtered clusters are of size greater than 3 and have a MaxSegNNDist of less than 0.5 (TrustFilter, boolean); the date edited (Date_Changed), and the clustering parameters used for each cluster (Rho_Threshold in meters, Theta_Threshold in degrees) and a unique identification calculated based on the start and endpoints (ClusterHash).

Data Set S2. Linked dike clusters for the Deccan Traps including the four identified subswarms: Saurashtra, Narmada-Tapi, Central and Coastal. Due to their overlap Central and Coastal Swarms have been combined in this dataset into the Central Swarm. This dataset was produced using the Agglomerative Clustering algorithms using the parameters set in Table 1. This dataset is in the format of a CSV file but can be read into GIS programs using Well Known Text (WKT) linestring. This dataset uses the a WGS 84 projection (EPSG:3857). The file includes the start and end points of the average line in the cluster and it's mid points, cluster length and width (Xstart, Xend, Xmid, Ymid, in meters and UTM coordinates, Dike Cluster Width or R_Width, Dike Cluster Length or R_Length all in meters); calculated average ρ and θ for the Hough Transform ρ units measured in meters, θ units measured in degrees, unless otherwise stated); the origin used for the Hough Transform which is different for each subswarm (xc,yc, meters in UTM coordinates); average slope and intercept (AvgSlope, AvgIntercept meters);

range and standard deviation for ρ and θ for all objects in the cluster (ρ units measured in meters, θ units measured in degrees); cluster size (Size); sum of segment lengths in a cluster (SegmentLSum, meters); whether the cluster crosses between negative and positive values (ClusterCrossesZero, boolean); overlap as calculated in the main text where the length of overlap is normalized by the sum of segment lengths in a cluster; maximum number of overlapping segments (nOverlapingSegments); twist angle which is the difference in angle betweeen the average cluster line and the average line formed by cluster midpoints (EnEchelonAngleDiff, degrees); the p-value for the midpoint line fit of the segments where $p < 0.05$ is considered to be a significant fit (EEPValue); the maximum, median, and minimum segment nearest neighbors distances in the cluster which is calculated using the cartesian midpoints of each segment and normalized by the Cluster Length (MaxSegNNDist, MedianSegNNDist, MinSegNNDist); characterization of each cluster as filtered or not, filtered clusters are of size greater than 3 and have a MaxSegNNDist of less than 0.5 (TrustFilter, boolean); the date edited (Date_Changed), and the clustering parameters used for each cluster (Rho_Threshold in meters, Theta_Threshold in degrees) and a unique identification calculated based on the start and endpoints (ClusterHash).

Data Set S3. Dike segment data for Spanish Peaks and Dike Mountain located in the Rio Grande Rift of Colorado. This dataset was digitized using QGIS based on the map by (Johnson, 1961). This dataset is in the format of a CSV file but can be read into GIS programs using Well Known Text (WKT) linestring. This dataset uses the a UTM Zone13N projection (EPSG:32613). The file includes the start, end points, and midpoints of the dikes; segment length; calculated ρ and θ for the Hough Transform; the origin used for the Hough Transform which is different for each subswarm (xc,yc); dike rock type if known; and a unique identification calculated based on the start and endpoints. This dataset has been preprocessed to remove curving dikes and is the data set used to produce later products (Data set S6).

Data Set S4. Linked dike clusters for the Spanish Peaks and Dike Mountain. This dataset was produced using the Agglomerative Clustering algorithms using the parameters set in Table 1. This dataset is in the format of a CSV file but can be read into GIS programs using Well Known Text (WKT) linestring. This dataset uses the a UTM Zone 13N projection (EPSG:32613). The file includes the start and end points of the average line in the cluster and it's mid points, cluster length and width (Xstart, Xend, Xmid, Ymid, in meters and UTM coordinates, Dike Cluster Width or R_Width, Dike Cluster Length or R_Length all in meters); calculated average ρ and θ for the Hough Transform ρ units measured in meters, θ units measured in degrees, unless otherwise stated); the origin used for the Hough Transform which is different for each subswarm (xc,yc, meters in UTM coordinates); average slope and intercept (AvgSlope, AvgIntercept meters); range and standard deviation for ρ and θ for all objects in the cluster (ρ units measured in meters, θ units measured in degrees); cluster size (Size); sum of segment lengths in a cluster (SegmentLSum, meters); whether the cluster crosses between negative and positive values (ClusterCrossesZero, boolean); overlap as calculated in the main text where the length of overlap is normalized by the sum of segment lengths in a cluster; maximum number of overlapping segments (nOverlapingSegments); twist angle which is the difference in angle betweeen the average cluster line and the average line formed by cluster midpoints (EnEchelonAngleDiff, degrees); the p-value for the midpoint line fit of the segments where $p < 0.05$ is considered to be a significant fit (EEPValue); the maximum, median, and minimum segment nearest neighbors distances in the cluster which is calculated using the cartesian midpoints of each segment and normalized by the Cluster Length (MaxSegNNDist, MedianSegNNDist, MinSegNNDist); characterization of each cluster as filtered or not, filtered clusters are of size greater than 3 and have a MaxSegNNDist of less than 0.5 (TrustFilter, boolean); the date edited (Date_Changed), and the clustering parameters used for each cluster (Rho_Threshold in meters, Theta_Threshold in degrees) and a unique identification calculated based on the start and endpoints (ClusterHash).

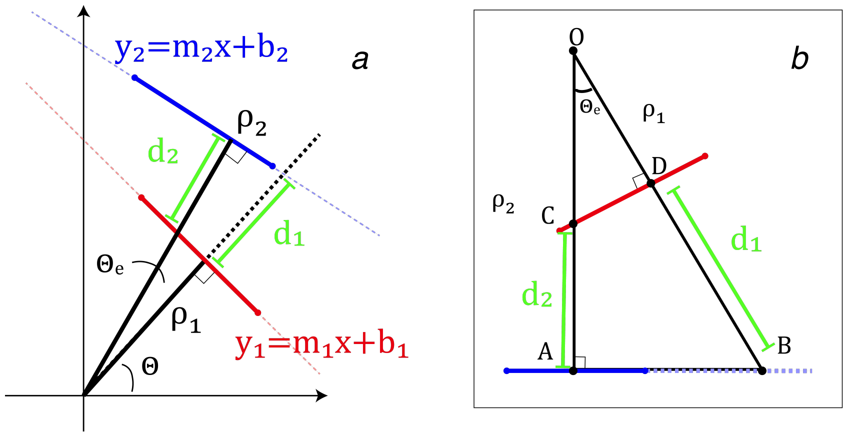

Figure S1. Figure S1 : **A.** shows two line in Cartesian space that are subparallel and the distances (d_1 and d_2) between them relative to the Hough transform ρ distances. **B.** shows the same view with the origin at the top and intersections labeled A, B, C, D.

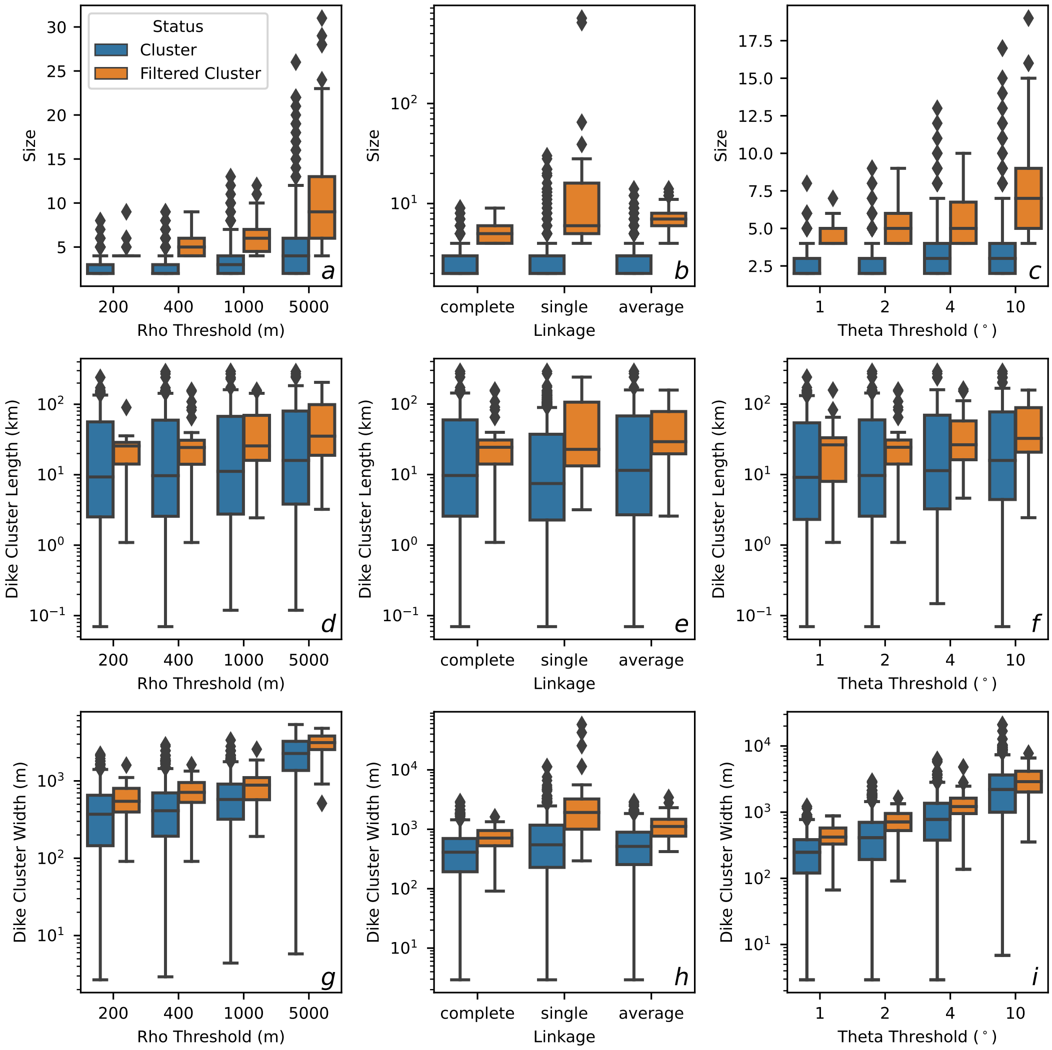

Figure S2. Figure S2 : **A.-C.** show a box and whisker plot of cluster size for the whole linked dataset and filtered clusters when changing ρ threshold, linkage, and θ threshold respectively. **D.-F.** show a box and whisker plots of cluster length in log scale. **G.-I.** show a box and whisker plots of cluster width in log scale.

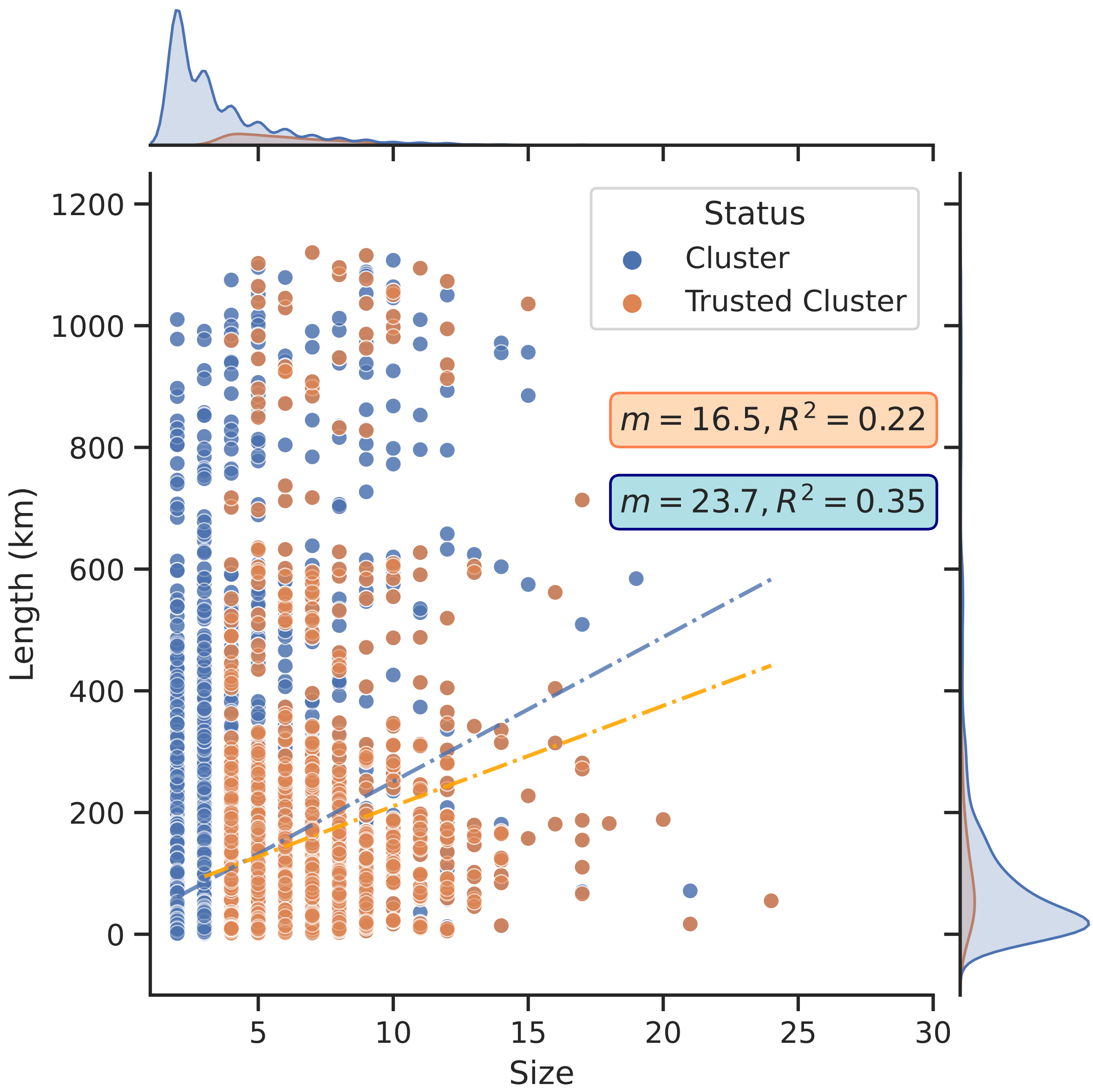

Figure S3. Cluster length as a function of cluster size for all dike segment datasets with all clusters in blue and filtered clusters in orange. Overlaid are two regression curves (dotted dashed lines) and their respective fits. Overall neither the full cluster database nor the filtered database show strong relationship between cluster size and cluster length. This is to be expected since clustering is done in the Hough transform space irrespective of Cartesian location. Very long dike clusters ($> 200km$) are seen with cluster sizes of 2-18 although clusters of over 5 are relatively rare.

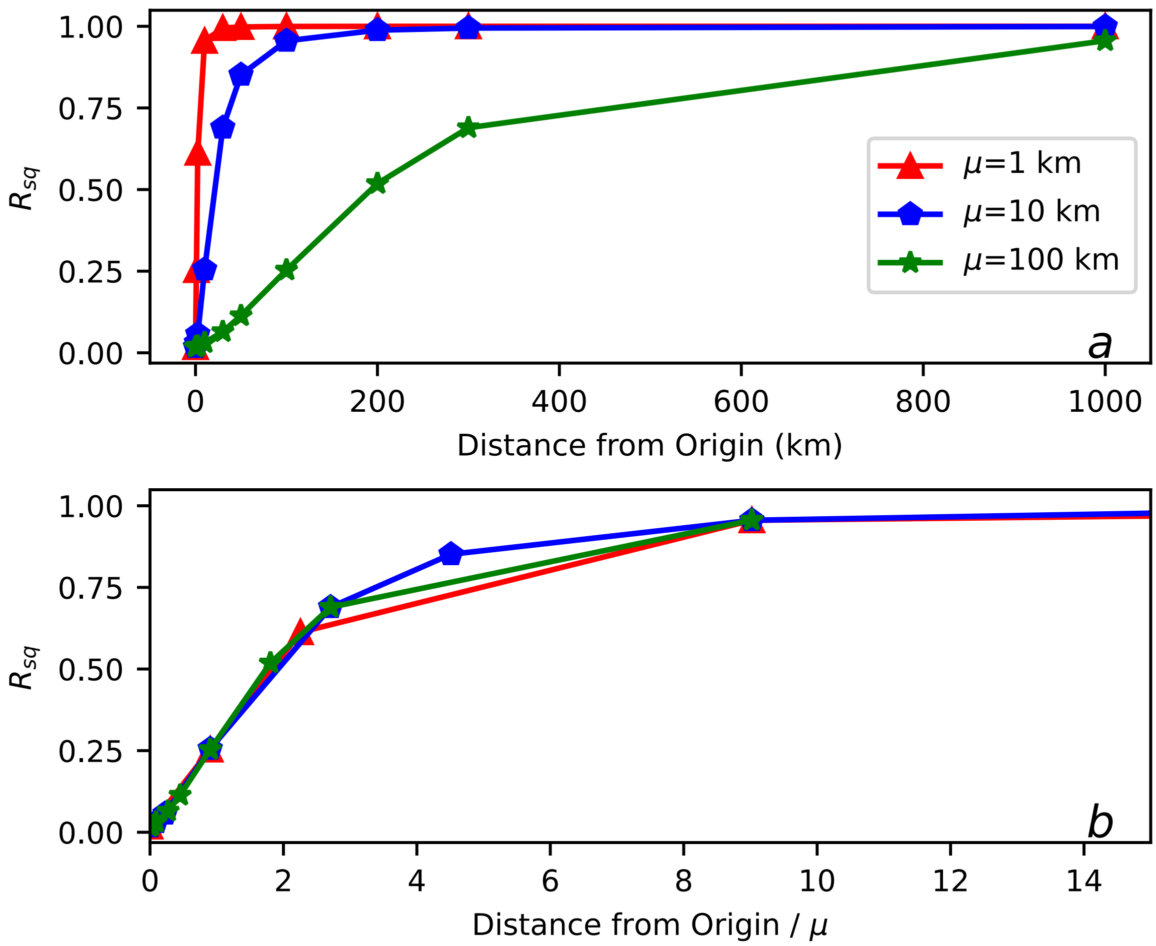

Figure S4. **A** For three synthetic randomly oriented swarms of different scales, the distance from the Hough Transform Origin is plotted against the goodness of fit for a radial swarm. **B** shows the same but with distance normalized by the scale of the swarm (μ), the standard deviation of ρ.

References

Everitt, B. (1980). *Cluster analysis* (Vol. 14) (No. 1). doi: 10.1007/BF00154794

Johnson, R. (1961). Patterns and Origin of Radial Dike Swarms Associated with West Spanish Peak and Dike Mountain, South-Central Colorado. *Geological Society of America Bulletin*, *72*(April), 579–590.

Morriss, M. C., Karlstrom, L., Nasholds, M. W., & Wolff, J. A. (2020). The chief Joseph dike swarm of the Columbia river flood basalts, and the legacy data set of William H. Taubeneck. *Geosphere*, *16*(4), 1793–1817. doi: 10.1130/GES02173.1

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... Contributors, S. . (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272.