Evaluating the accuracy of binary classifiers for geomorphic applications

Matthew W. Rossi[1]

[1]Earth Lab, Cooperative Institute for Research in Environmental Sciences, The University of Colorado, Boulder, CO 80303, USA

Correspondence to: Matthew W. Rossi (matthew.rossi@colorado.edu)

Abstract. Increased access to high resolution topography has revolutionized our ability to map out fine-scale topographic features at watershed- to landscape-scales. As our 'vision' of land surface has improved, so has the need for more robust quantification of the accuracy of the geomorphic maps we derive from these data. One broad class of mapping challenges is that of binary classification where remote sensing data are used to identify the presence or absence of a given feature. Fortunately, there are a large suite of metrics developed in the data sciences that are well suited to quantifying pixel-level accuracy of binary classifiers. This analysis focuses on how these metrics perform when there is a need to quantify how the number and extent of landforms are expected to vary as a function of the environmental forcing (e.g., due to climate, ecology, material property, erosion rate). Results from a suite of synthetic surfaces show how the most widely used pixel-level accuracy metric, *F1-score*, is particularly poorly suited to quantifying accuracy for this kind of application. Well-known biases to imbalanced data are exacerbated by methodological strategies that calibrate and validate classifiers across settings where feature abundances vary. *Matthews Correlation Coefficient* largely removes this bias over a wide range of feature abundances, such that the sensitivity of accuracy scores to geomorphic setting instead embeds information about the size and shape of features and the type of error. If error is random, *Matthews Correlation Coefficient* is insensitive to feature size and shape, though preferential modification of the dominant class can limit the domain over which scores can be compared. If the error is systematic (e.g., due to co-registration error between remote sensing datasets), this metric shows strong sensitivity to feature size and shape such that smaller features with more complex boundaries induce more classification error. Future studies should build on this analysis by interrogating how pixel-level accuracy metrics respond to different kinds of feature distributions indicative of different types of surface processes.

1 Motivation

High resolution topographic datasets are transforming our ability to characterize the fine-scale structure of the Earth's surface (Passalacqua et al., 2015). Airborne lidar especially, has changed how geomorphic fieldwork is conducted by enabling scientists to quantify the form and extent of meter-scale features over large areas (Roering et al, 2013). Because lidar 'sees' through vegetation, lidar has accelerated progress in both discovery science and testing hypotheses where the prevalence of features is expected to vary as a function of the environmental forcing (e.g., in response to differences in climate, ecology, material property, erosion rate). Airborne lidar has now been used to map mima mounds (Reed & Amundson, 2012), termite mounds (Levick et al., 2010; Davies et al., 2014), , tree throw pits and mounds (Roering et al., 2010; Doane et al., 2023), landslide boundaries and classes (Jaboyedoff et al, 2012; Bunn et al., 2019; Prakesh et al., 2020), channel networks (Pirotti & Tarolli, 2010; Clubb et al., 2014; Korzeniowska et al., 2018), bedrock structure (Cunningham et al., 2006; Pavlis and Bruhn, 2011; Morell et al., 2017), and bedrock exposure (DiBiase et al., 2012; Milodowski et al., 2015; Rossi et al., 2020).

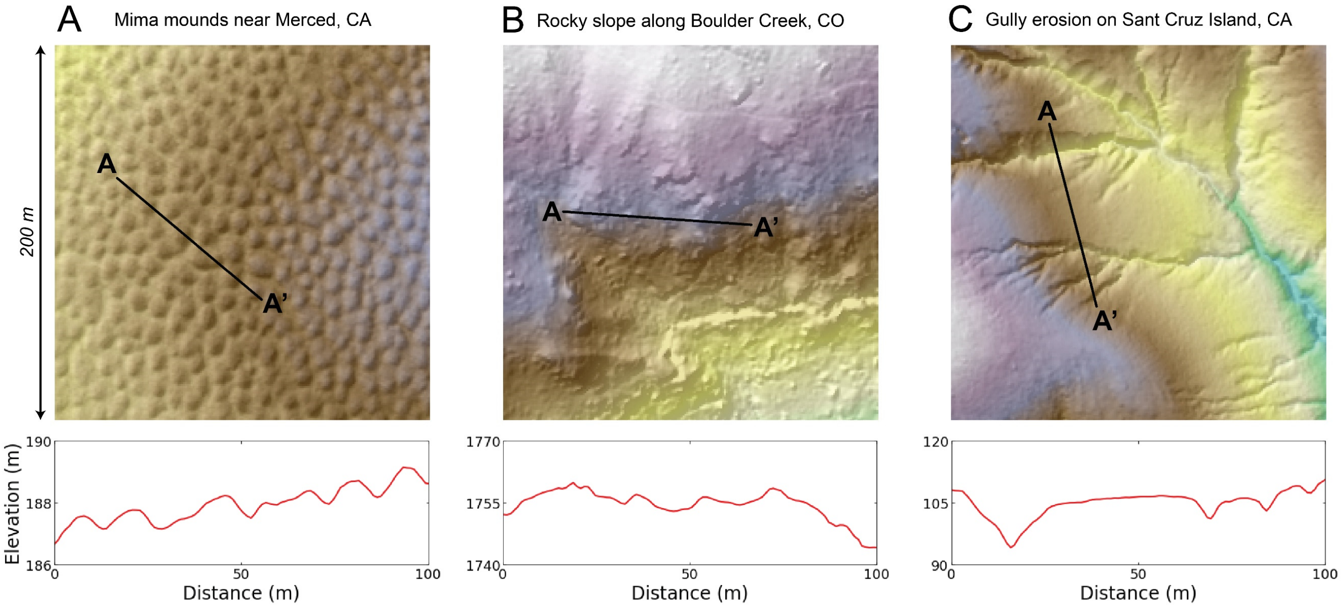

Figure 1: (A) Mima mounds near Merced, CA, USA, (B) bedrock outcrops along Boulder Creek, CO, USA, and (C) gully erosion on Santa Cruz Island, CA, USA as observed from 1-m shaded relief maps. Note that even though the areal extent is the same among these scenes (200 x 200 m), topographic relief is drastically different (total relief in A is 7 m, in B is 146 m, and in C is 76 m). 100-m elevation transects from A to A' for each site are shown to illustrate how different features manifest as roughness elements in the topography. Airborne lidar for the mima mound and rocky slope sites was flown by the National Center for Airborne Laser Mapping (NCALM). Airborne lidar for the gully erosion site was flown by the United States Geological Survey (USGS). All lidar datasets were downloaded from OpenTopography (Reed, 2006; Anderson et al., 2011; 2010 Channel Islands Lidar Collection, 2012). Interpretations of features classified from lidar data can be found in Reed & Amundson (2011), Rossi et al. (2020), and Korzeniowska et al. (2018) for the mima mound, rocky slope, and gully sites, respectively.

Figure 1 shows three examples of features that can be mapped using 1-m airborne lidar data. The utility of binary classification of feature locations for each of these geomorphic applications is clear. However, examples also highlight how the number, size, shape, amplitude, and pattern of features can vary. Regular, repeating morphologies with a characteristic spatial scale (e.g., mima mounds in Fig. 1A; Reed and Amundson, 2011) pose different challenges to classification than irregular, heterogeneous morphologies that occur at many scales (e.g., bedrock exposure in Fig. 1B; Rossi et al., 2020). Furthermore, the importance of flowing water on surface processes means that many geomorphic features form directional networks with substantial anisotropy (e.g., gully erosion in Fig. 1C; Korzeniowska et al., 2018). Perhaps unsurprisingly then, accuracy assessment in the geomorphic literature has varied a lot even as formal methods for evaluating pixel-level accuracy of binary classifiers are now becoming standard practice in the remote sensing and machine learning literature (e.g., Wang et al., 2019; Prakesh et al., 2020; Agren et al, 2021). Slow adoption of these standard methods in accuracy assessment may arise from two tendencies of geomorphic studies that employ lidar classifiers: 1. Process-based studies are typically more interested in the *properties* and *densities* of features rather than their contingent locations; 2. Classifiers are expected to work across *large gradients* in the prevalence of features to test our understanding of the relevant transport laws at play. The former tendency arises from the fact that predicting the actual locations of features (e.g., mounds, outcrops, channels) is not usually a viable target for numerical models of landscapes where uncertainty in initial conditions and the stochastic nature of processes preclude a deterministic forecasting of finer-scale locations of features (e.g., Barnhart et al., 2020). The latter tendency arises from the need to use classified data to constrain natural experiments where geomorphic transport laws (Dietrich et al., 2003) can be tested against governing variables (e.g., across climo-, eco-, litho-, or tectono-sequences). As shown below, these tendencies can be at odds with pixel-level accuracy metrics designed to assess positional accuracy for balanced data (i.e., data where the frequency of positive and negative values are similar).

Nevertheless, there are several important benefits to adopting pixel-level accuracy metrics when reporting the success of geomorphic classifiers. First, these metrics provide common standards for evaluating classifier accuracy across studies, including direct comparison between proxy-based classifiers and those developed using machine learning. Second, trends in pixel-level accuracy scores may reveal patterns in the spatial structure of error. Third, pixel-level measures are easy to apply to new objectives as long as their limitations are properly considered. This paper focuses on how two widely used metrics, *F-measures* (van Rijsbergen, 1974; Chinchor, 1992) and *Matthews Correlation Coefficient* (Matthews, 1975; Baldi et al., 2000), perform when the research design intentionally calibrates and tests binary classifiers across large gradients in how balanced the data are. The general approach is to synthetically generate 'model' and 'truth' data that have a known error structure. Pixel-level accuracy scores are then calculated as a function of feature abundance. Despite the simplicity of the scenarios considered, this analysis helps constrain the range over which pixel-level metrics can be reliably compared across gradients in feature abundance. Synthetic scenarios also reveal how the shape and scale of individual objects can strongly influence pixel-level scores when there are small co-registration errors between model and truth data.

2 Approach

One common use of binary classifiers is to build an inventory of feature boundaries and abundances using remotely sensed data. This typically entails using scenes where 'truth' is known through detailed field or air photo mapping. An algorithm built from an independent data source (e.g., lidar) is then used to 'model' the locations of features. Models are commonly trained and tested so that the classifier can be used for larger scale geomorphic mapping. If the density, size distribution, and form of features varies from scene to scene, then it is important to understand how pixel-level accuracy metrics will perform as a function of scene-level properties (e.g., feature fraction). To mimic this task, this paper examines how two widely used accuracy metrics, *F1-score* and *Matthews Correlation Coefficient* (*MCC*), behave on synthetic truth and model data. Synthetic truth data is generated by randomly placing features in a scene at a given abundance. Model data is either independent from truth data or derived from the truth data using an assumed error structure. Pixel-level accuracy scores are then calculated for each scene.

2.1 Grid generation

To generate 'truth' grids of features within a matrix, the pseudo-random number generator in NumPy is used to create a scene of size m x n cells. Continuous values are converted into binary classes (0 = matrix; 1 = feature) based on a user-specified value for the feature fraction (f_f), which is simply the fraction of the surface covered by features. The simplest scenario is for features with a size of one pixel. While synthetic surfaces are scale free, results are reported assuming a grid spacing of 1-m to represent a typical case using airborne lidar. To simulate features that have a scale greater than one square meter, the pseudo-random numbers instead specify a first guess at the locations of the centres of incipient features. The first guess at the number of features is calculated by finding the integer number of features of length, l, that most closely matches f_f. However, as the number of feature centres increases, so does the probability that two neighbouring objects overlap and coalesce into a larger object. As such, the first guess generally produces an actual feature fraction lower than the user-specified value. The ratio between the specified f_f and this underestimate is then used to proportionally increase the number of incipient features in the model domain. The process is iterated until either the synthetic fraction is within 0.5% of the specified value or fifty iterations, whichever comes first. The number of incipient objects is always higher than the actual number of objects in the scene because smaller incipient features increasingly coalesce into larger objects at higher feature fractions.

All scenarios in this study rely on comparing simulated 'truth' and 'model' grids across the full range of feature fractions ($0 < f_f < 1$). Where the truth and model data are independent of each other, the two grids are generated using different pseudo-random seed numbers in NumPy (section 3). In scenarios where the model grid is dependent on the truth grid, the model grid is a copy of the truth data using the specified error structure. Details for how random error (section 4.1), systematic error (section 4.2), and random plus systematic error (section 4.3) are implemented are described in context below. For each scenario, the truth and model grids are evaluated by building the confusion matrix and calculating accuracy metrics at each feature fraction (section 2.2).

2.2 Pixel-level accuracy metrics

While there are many metrics used to quantify the accuracy of binary classifiers, the focus of this paper is on two of the most widely used ones: the *F1-score* and *Matthews Correlation Coefficient* (*MCC*). These metrics are frequently used to evaluate pixel-level performance of classified maps generated from machine learning (e.g., Wang et al., 2019; Prakesh et al., 2020; Agren et al, 2021). Application of these metrics need not be limited to the training and testing of machine learning algorithms. They are broadly useful to any binary classification task where positional accuracy is important. Both *F1-score* and *MCC* can be calculated directly from the confusion matrix. The confusion matrix for binary classification is a 2x2 table where the column headers are the true classes and the row headers are the model classes, thereby summarizing the occurrence of the four possible classification outcomes: True Negatives (TN), True Positives (TP), False Positives (FP), and False Negatives (FN).

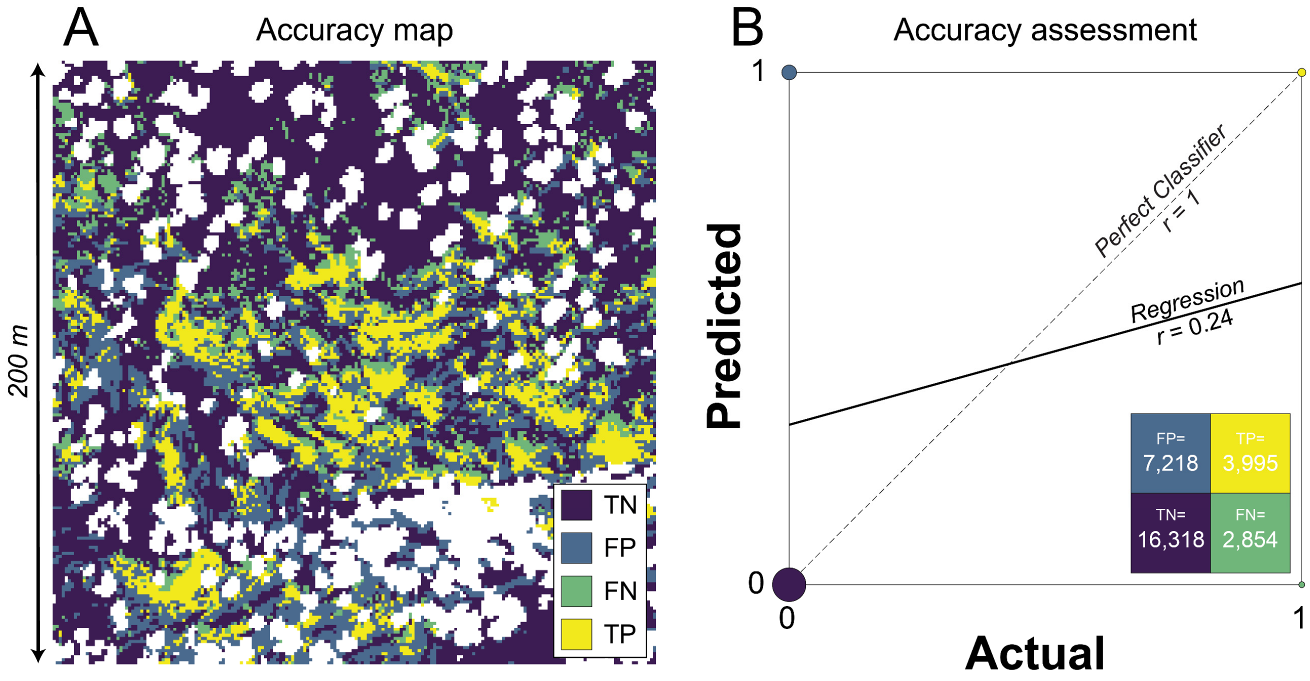

Figure 2: (A) Pixel classes for Fig. 1B and (B) the corresponding confusion matrix (inset) and correlation plot (main). In A, the four outcomes of the binary classification are shown in colour [TN = True Negatives; FP = False Positives; FN = False Negatives; TP = True Positives]. The areas in white were obscured by the vegetation canopy in air photos (24% of area) and thus excluded from accuracy assessment. In B, the colours of each cell in the confusion matrix and each point in the plot are the same as in A. The number of observations for each class is shown in the confusion matrix and point sizes on the plot are scaled to the relative frequency of each value. This classified map is site P01 from Rossi et al. (2020), where more details on mapping methods are described.

For example, the example scene in Figure 1B is readily reclassified into these four outcomes (Fig. 2A) using the feature mapping from Rossi et al. (2020). The frequency of these outcomes is summarized using the confusion matrix (Fig. 2B inset). The simplest accuracy metric is the overall accuracy (OA), and its complement the error rate (ER), where:

$$OA = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$ER = \frac{FP+FN}{TP+TN+FP+FN} \tag{2}$$

While *OA* and *ER* are straightforward to calculate, they provide little insight into the relative frequencies of FP and FN. To address this limitation, there are a large family of accuracy metrics that better characterize different types of error. For example, *precision* and *recall* characterize the relative frequencies of FP and FN explicitly. *Precision*, also known as the positive predictive value, is the ratio of true positives to all positives predicted by the model (accounts for FP):

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

Recall, also known as the true positive rate, is the ratio of true positives to all positives (accounts for FN):

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

Figure 2 is an example where the *precision* is low (0.36), but the *recall* is reasonably good (0.58) (Table 1). *F-measures* were designed to summarize *precision* and *recall* into a single metric (van Rijsbergen, 1974; Chinchor, 1992). The case where both are equally weighted is referred to as the *F1-score*, where:

$$F1\text{-}score = \frac{2\times TP}{(2\times TP)+FP+FN} \tag{5}$$

By representing the harmonic mean of *precision* and *recall*, this metric accounts for both errors of omission and commission. *F1-scores* only characterize the success at identifying the target class, and low values can occur even if the overall accuracy is high because it excludes True Negatives. As such, this metric is sensitive to the prevalence of positive values. Higher *F1-scores* are favoured when the positive class is more abundant (e.g., Chicco and Jurman, 2020). Related to this sensitivity to imbalanced data is the property of asymmetry. Asymmetric metrics are those where the accuracy score differs when the target classes are switched. Table 1 shows that the *F1-score* for Figure 2 would be 72% higher if the target feature was soil instead of bedrock. Asymmetry arises because there is more soil than bedrock in the scene and TN are not included in calculations of *precision*, *recall*, or *F1-score*. These well-known limitations of *F-measures* are better handled by metrics that incorporate all four classes of the confusion matrix. One such metric is *Matthews Correlation Coefficient* (*MCC*), where:

$$MCC = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \tag{6}$$

MCC is equivalent to a Pearson's correlation coefficient where the model classes are regressed against the true classes in a binary classification task (Fig. 2B). Values of *MCC* can be similarly interpreted where -1.0 indicates perfect anti-correlation, 0 is a random model, and 1.0 indicates perfect correlation. And while *MCC* is just one of several metrics that include all four quadrants of the confusion matrix (e.g., Balanced Accuracy, Markedness, Cohen's Kappa), recent work suggests that *MCC* is the most robust to imbalanced data (Chicco and Jurman, 2020; Chicco et al., 2021a; Chicco et al., 2021b). In this analysis, I report a normalized version of *MCC* as:

$$nMCC = \frac{MCC+1}{2} \quad (7)$$

By re-scaling *MCC* from zero to one, *nMCC* facilitates comparison with *F1-score* on plots and in discussion. It is worth noting though that interpretations of low values of *nMCC* differ from interpretations of low values of *F1-score*. The former implies anti-correlation between model and truth data while the latter does not. For example, the scene in Figure 2 indicates a weak positive correlation (i.e., *nMCC* greater than 0.5) even though the *F1-score* is lower than 0.5 (Table 1). As such, direct comparison of these metrics should be done with caution.

Table 1: Accuracy metrics for Figure 2 using the alternative target classes of bedrock and soil.

Target Class	***OA****	***ER****	***Precision***	***Recall***	***F1-score***	***MCC****	***nMCC****
Feature (bedrock)	0.67	0.33	0.36	0.58	0.44	0.24	0.62
Feature (soil)	0.67	0.33	0.85	0.69	0.76	0.24	0.62

* *Metrics that do not vary as a function of the target class in binary classification.*

3 Independence between truth and model data

The distinction between pixel-level and scene-level accuracy, in part, motivates the approach taken to examine how accuracy metrics handle imbalanced data in this study. Pixel-level accuracy requires that the precise locations of features are honoured, with a lower bound to feature detection set by the spatial resolution of the data used. Scene-level accuracy characterizes the mismatch between model and truth data at some coarser scale and typically assesses statistical properties of the target feature class (e.g., bedrock fraction, mound densities, drainage densities). While high pixel-level accuracy ensures high scene-level accuracy, the converse need not be true. Given the importance of developing binary classifiers that work across a range of feature densities and sizes, there is a need to better understand how pixel-level accuracy metrics perform across a range of scene-level properties like feature fraction. One mark of a good accuracy metric is its ability to diagnose the case of independence. In this context, independence means that the locations of features in the model contain no information about the true locations of features. If accuracy metrics produce similar scores when the model and truth data are independent from each other, then it means the metric can be reliably compared for different feature fractions. A perhaps trivial example is the case where feature fractions are assumed to be constant (e.g., total feature coverage) regardless of the true feature fraction. A more interesting example is the case where scene-level fractions are the same in the truth and model data (i.e., high scene-level accuracy) but where the actual locations of features are unrelated (i.e., low pixel-level accuracy).

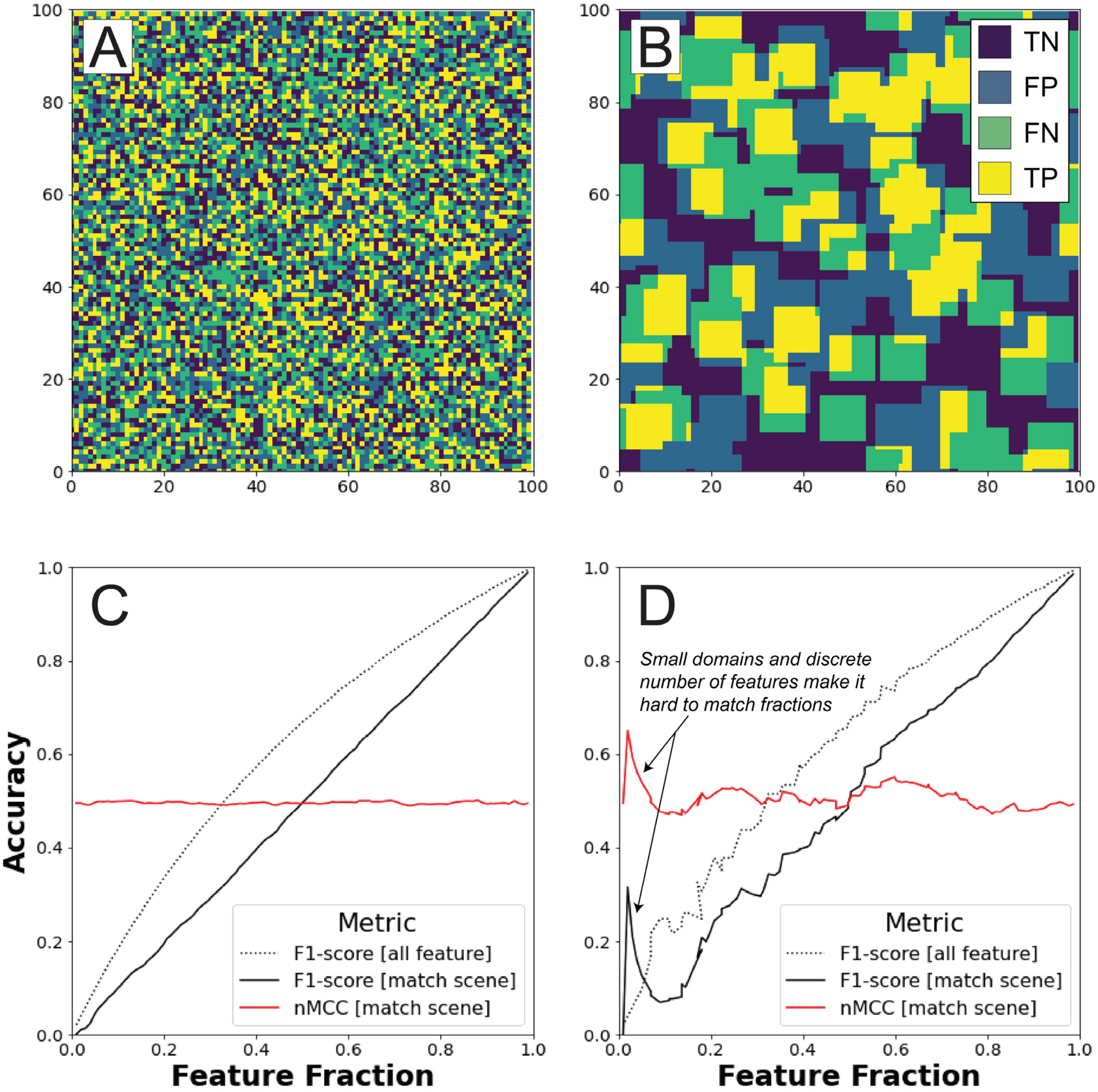

Figure 3: Classified 100 x 100 m maps of (A) 1-m and (B) 10-m long incipient features showing the four classification outcomes (*TN*: True Negatives, *FN*: False Negatives, *FP*: False Positives, *TP*: True Positives). How accuracy scores vary as a function of feature fraction are also shown for (C) 1-m and (D) 10-m long incipient features, respectively. The ‘all feature’ scenario is where the model assumes the entire surface is feature with no matrix, regardless of scene-level properties. The ‘match scene’ scenario is where the model data matches the actual feature fraction, but whose feature locations are independent of each other. In A-B, example maps are shown for the case where fifty percent of the surface is covered by features. In C-D, *normalized Matthews Correlation Coefficient* (*nMCC*) is only shown for the ‘match scene’ scenario because it is undefined in the ‘all feature’ scenario.

Figure 3 shows the sensitivity of *F1-score* and *nMCC* to imbalanced data when the model and truth data are independent from each other ($m = n = 100$). Each scenario assumes features are randomly distributed throughout the scene for any given feature fraction. In the first scenario, the classifier predicts that the feature is found everywhere regardless of the truth data (dashed lines). Because this 'all feature' model produces neither False Negatives nor True Negatives, *nMCC* is undefined in this scenario (see eqs. 6-7). *F1-score* nonlinearly improves with increasing feature fraction and approaches unity as the actual fraction nears the 'all-feature' model. In the second scenario, the classifier is forced to match the feature fraction in the truth grid, though the locations of features in the model are independent from the truth data (solid lines). This represents a worst-case scenario for a classifier that successfully models the scene-level fraction while also providing zero predictive value at the pixel level. The values of *nMCC* rightly diagnose independence between the model and truth data by showing zero correlation across the full range of feature fractions (*nMCC* ~ 0.5). In contrast, *F1-score* increases as a linear function of feature fraction. As this and subsequent examples show, *F1-score* embeds a spurious correlation with feature fraction, all other things being equal, because the number of True Negatives is ignored. In contrast, *nMCC* provides a robust metric to evaluate positional error for classifiers that have been calibrated to scene-level properties. While these relationships do not depend on incipient feature size, larger mapping areas are needed to adequately sample the statistics of feature locations when incipient features are large with respect to the area of the scene (Fig. 3D). The noisy relationships in Figure 3D largely reflect the inability to match the specified feature fraction using a discrete number of random features whose locations are set by the specific pseudo-random seed used. In fact, 49% of the grids generated for Figure 3D failed to meet the 0.5% tolerance of specified feature fractions after fifty iterations. For subsequent analyses, larger 1000 x 1000 m scenes are used to mitigate the effect of domain size on accuracy scores. For the larger domain, nearly all (>99%) the subsequent grid pairs meet the tolerance criterion before fifty iterations, which manifest as smoother curves in plots.

4 Error structure and accuracy

The previous section showed how *F1-score* and *nMCC* vary as a function of feature prevalence for classifiers that only honoured scene-level attributes (i.e., feature fraction) with no predictive skill at identifying feature locations. While a useful baseline scenario, a good classifier should identify both the locations of features and reproduce scene-level attributes, albeit with some residual error. To illustrate these more realistic conditions, three different error scenarios are presented where the error structure is either random (section 4.1), systematic (section 4.2), or both (section 4.3). While actual sources of error in geomorphic studies are typically more complex, these simple scenarios facilitate interpretation and provide insight into how pixel-level accuracy scores perform when the research design explicitly samples across a gradient in feature prevalence.

4.1 Random error

The first error scenario considered is the situation where the binary classifier successfully identifies feature locations with a fixed rate of random error ($\bar{e}_r$). To create synthetic surfaces of this type, a truth grid is first generated (for Fig. 4 $m = n = 1{,}000$) for a given feature fraction. Features are assumed to occupy a single pixel, though results are robust to different sizes of incipient features because where error occurs is independent of feature locations. To produce the associated model grid, an error grid is first generated using a different pseudo-random seed than that used to generate the truth data. The continuous values of the error grid are converted to binary classes (0 = no error; 1 = error) using the specified error rate as the threshold. The error grid is then used to construct the model grid from the truth grid by flipping feature classifications wherever the error grid value equals one. Note that the maximum error rate shown in Figure 4 is fifty percent. This is the scenario where the truth and model data are least correlated. Increasing the error rate further will produce increasingly stronger negative correlations between the model and truth data. Once both truth and model grids are generated, *F1-score* and *nMCC* are calculated. This analysis is done for feature fractions that range from 0.01 to 0.99 and error rates from 5 to 50 percent.

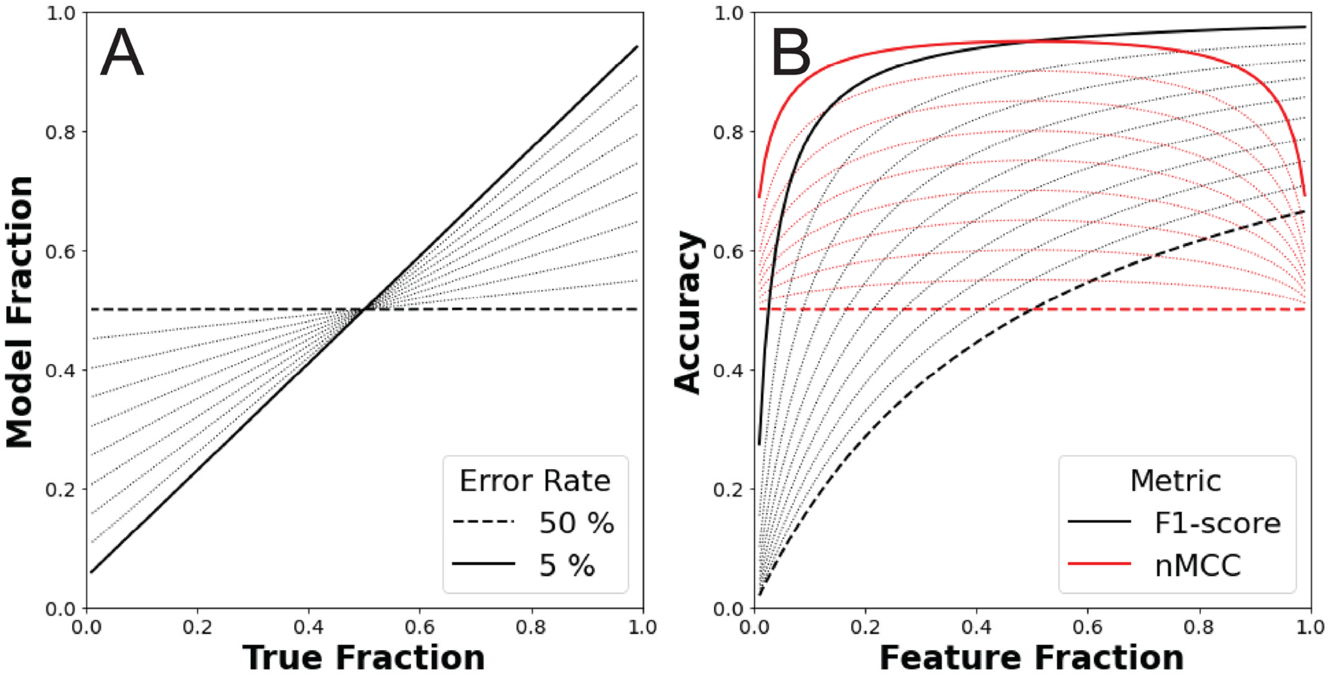

Figure 4: (A) Model feature fractions and (B) associated accuracy scores as a function of the true feature fraction in the random error scenario (1000 x 1000-m map area). In both plots, the minimum and maximum error rates are highlighted, and 5% increments of error rate are shown as dotted lines. In A, matching the model fraction to the actual fraction of bedrock is not enforced like in other scenarios (Figs. 3, 5). However, the two fractions are linearly related, and the slope of the relationship is directly related to the error rate (Appendix A). In B, lower rates of random error amplify the nonlinearity between *F1-score* and feature fraction while *nMCC* more uniformly improves across a broad range of feature fractions.

Figure 4 shows the results of this analysis for ten numerically simulated error rates. Results can be derived analytically from eqs. 5-7 and the imposed random error rate (Appendix A). However, presenting the results from numerical surfaces: 1. Ensures that synthetic scenes adequately sample population statistics; and 2. Facilitates integration with scenarios that include non-random error (section 4.3). As should be expected, Figure 4 shows that accuracy scores increase as error rates go down. However, the sensitivity of these scores is not uniform with respect to feature fraction. Much like in the previous scenario (Fig. 3), *F1-scores* always monotonically improve with increasing feature fraction. Note here though that the worst random error case (Fig. 4 dashed black line; 50% error rate) is not equivalent to the case where the model is independent from the truth data (i.e., the solid black line in Fig. 3). In the random error scenario, model data are correlated with, but not equal to, actual feature fractions (Fig. 4A). The fixed error rate preferentially modifies the larger frequency class near the endmember cases of zero and full coverage of the surface by features. This is most easily envisioned at the limits of feature abundance. If the actual surface is all features, then the random error model will produce matrix pixels in proportion to the error rate. Similarly, if the actual surface is all matrix, then the random error model will produce feature pixels in proportion to the error rate. For this error scenario, the slope of the relationship between modelled and actual feature fractions equals $1 - 2\bar{e}_r$ (Appendix A). The symmetry of the sensitivity of *nMCC* to a uniform, random error rate allows for comparison of map accuracies across a wide range of feature abundances, specifically over the domain over which *nMCC* is approximately invariant (Fig. 4B). In contrast, disentangling the spurious correlation between *F1-score* and feature fraction interacts with the preferential modification of classes in a complex way, leading to increasing nonlinearity for better classifiers with lower error rates.

4.2 Systematic error

The second error scenario considered is the situation where the binary classifier successfully identifies features with some imposed systematic error. This scenario is motivated by the common challenge of aligning two datasets collected using different sensors or collected at different times (e.g., Bertin et al., 2022). To create synthetic surfaces of this type, a truth grid is first generated (for Fig. 5 $m = n = 1{,}000$) for a given feature fraction and incipient feature size. Incipient features are randomly distributed throughout the model domain. To produce the associated model grid, a copy of the truth grid is linearly offset by one pixel to the right in the x-direction, though results are insensitive to the direction of the shift. By using wrap-around boundaries, synthetic truth and model grids always have identical feature fractions. Note that the systematic error rate ($\bar{e}_s$) is not constant and is instead a function of the feature fraction, the magnitude of the systematic offset, and the shape and size of features. Once both truth and model grids are generated, *F1-score* and *nMCC* are calculated. This analysis is done for feature fractions that range from 0.01 to 0.99 and for incipient feature sizes that range from 1x1 m to 10x10 m squares (i.e., areas of 1 to 100 pixels).

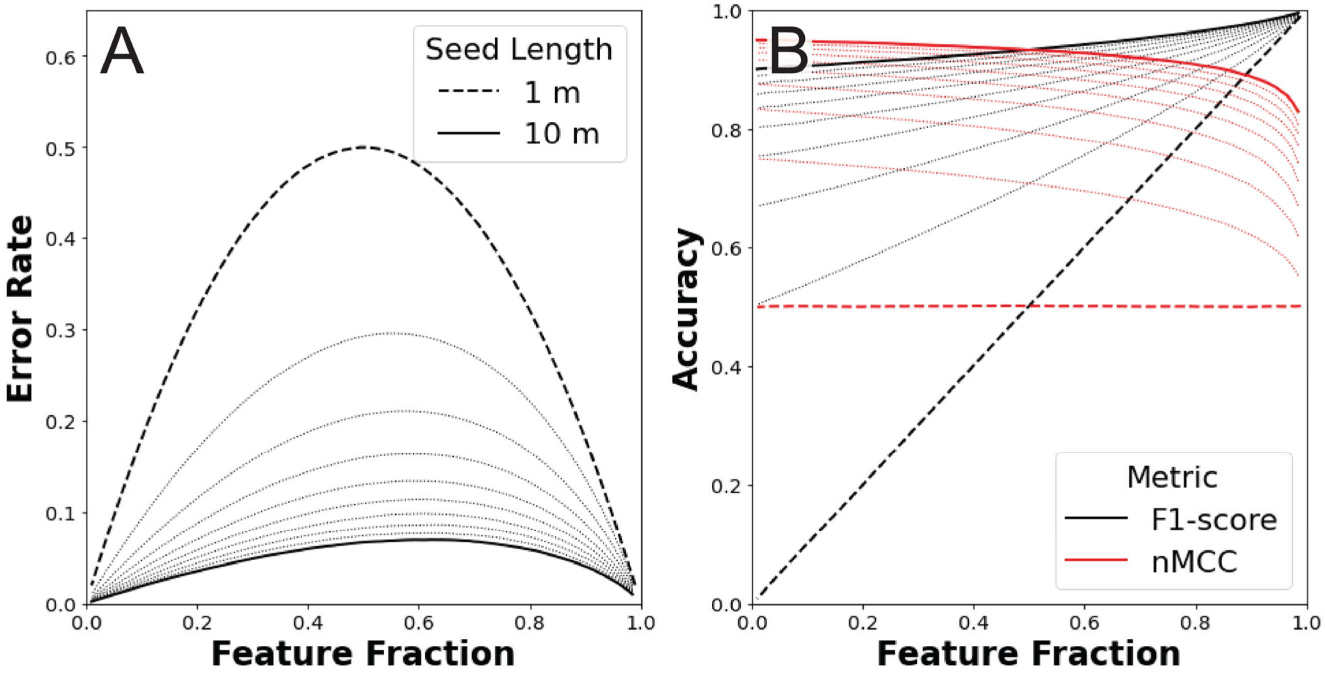

Figure 5: (A) Variable error rates and (B) associated accuracy scores as a function of the true feature fraction for the systematic error scenario (1000 x 1000 m map areas). In both plots, the minimum and maximum incipient feature lengths are highlighted, and 1-m increments are shown as dotted lines. In A, the error rate (eq. 2) is non-uniform with lower rates at both low and high feature fractions. As incipient feature size gets larger, the error rate function becomes increasingly asymmetrical with peak values at 0.5 and 0.66 bedrock for 1- and 10-m long seeds, respectively. In B, the non-uniform error rates lead to more linear relationships between *F1-score* and feature fraction than in the case of random error (Fig. 4B). In contrast, *nMCC* shows modest negative relationships with feature fraction for all incipient feature sizes.

Figure 5 shows the results of this analysis for ten different incipient seeds that span from 1 to 10 m in length (1 to 100 m^2). While results throughout this paper are discussed in terms of a scale typical to airborne lidar (i.e., 1-m spatial resolution), the relationships shown here are better cast as the ratio of the incipient feature scale (i.e., seed length in pixels) to the error scale (1 pixel length) where the feature detection limit is one pixel. When systematic error is of order feature length, systematic error mimics the case where the truth and model data are independent (e.g., compare long dashed lines in Fig. 5B to solid lines in Fig. 3C-D). As the systematic error gets small with respect to the incipient feature size, both *F1-score* and *nMCC* improve. The largest improvements occur for small incipient feature sizes and at low feature fractions (Fig. 5B). When feature fractions are low, the error is largely due to the geometric effect of the shift of individual square objects surrounded by matrix. As feature fraction increases, incipient objects increasingly coalesce into a smaller number of objects, and the error is set by these more complex geometries (see discussion in section 5.2). Figure 5A shows that increasing the incipient feature size leads to lower error rates and increasing asymmetry in the error rate function, where the highest error is biased towards higher feature abundances. These error rate functions manifest as a modest negative relationship between *nMCC* and feature fraction regardless of incipient feature size (Fig. 5B). The asymmetric error structure also impacts *F1-score*, albeit in a way that is much harder to diagnose due to the spurious correlation between *F1-score* and feature fraction (Figs. 3-4). The notion of systematic error in scene-level mapping was envisioned for situations where co-registration error between the remote sensing data used to map 'truth' and the remote sensing data used to build the classifier produce a systematic, translational offset. Strictly speaking then, this synthetic scenario represents the case where a translational offset is the same for all scenes, a plausible situation if the truth and model data for different scenes were acquired at the same time and in the same way. However, even under the less stringent condition where co-registration errors are oriented differently in different scenes (i.e., due to different acquisition parameters and times), the relationships shown in Figure 5 will still hold as long as the magnitude of the systematic error is similar across scenes and there is no preferred orientation to feature objects.

4.3 Random plus systematic error

The third error scenario considered is the situation where the binary classifier is systematically offset from the truth grid with an additional random error term. To create synthetic surfaces of this type, a truth grid is first generated (for Fig. 6 m = n = 1,000) for a given feature fraction and incipient feature size. Incipient features are randomly distributed throughout the model domain. To produce the associated model grid, a copy of the truth grid is first linearly offset by one pixel to the right in the x-direction, using a wrap-around boundary condition. A random error grid is then generated using a different pseudo-random seed than that used to generate the truth data. The continuous values of the error grid are converted to binary classes (0 = no error; 1 = error) using a random error rate of 0.05 as a threshold. The error grid is used to flip classifications in the offset feature grid wherever the error grid value equals one. Note that feature fractions in the model need not match the truth data, and error rates are now a function of the feature fraction, the magnitude of the systematic offset, the size and shape of features, and the random error rate. Once both truth and model grids are generated, *F1-score* and *nMCC* are calculated. This analysis is done for feature fractions that range from 0.01 to 0.99 and incipient feature sizes that range from 1x1 m to 10x10 m squares (e.g., areas of 1 to 100 pixels).

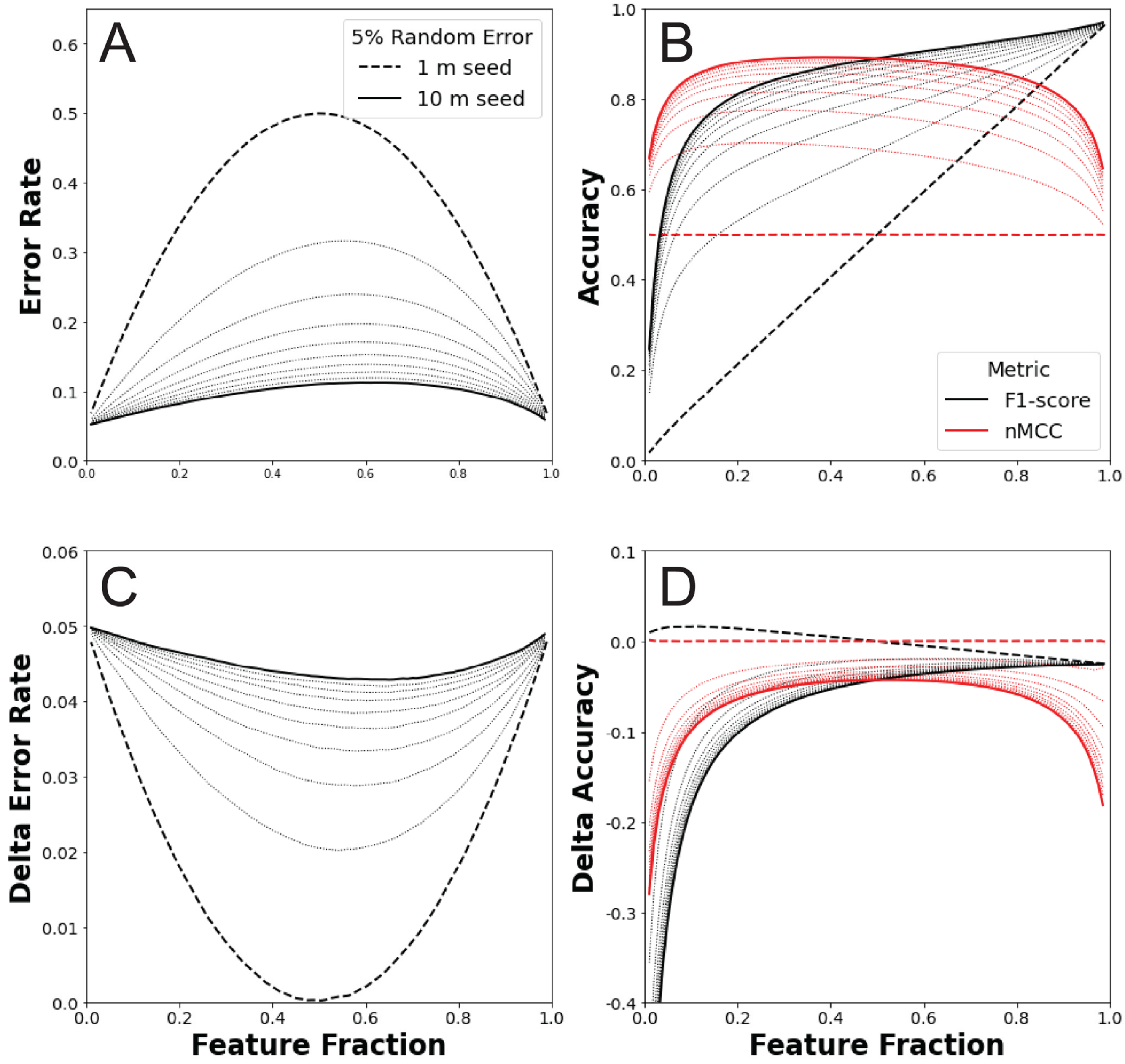

Figure 6: (A) Variable error rates and (B) associated accuracy scores as a function of the true feature fraction for the systematic plus random error scenario (1000 x 1000-m map areas). These panels are analogous to Figure 5A and 5B but now include a 5% random error term. Differences in (C) error rates and (D) accuracy scores between this scenario and systematic error alone (Fig. 5) are shown to enable comparison. In C, the additional 5% random error term is linearly added to the systematic error term at the endmember cases of zero and total feature coverage. The random error translates into something less than 5% for intermediate cases with minima near zero for 1-m seeds and 0.043 for 10-m seeds. In D, *nMCC* exhibits strong reductions from systematic error alone near endmember cases (high negative values) and a muted, more uniform reduction at intermediate values. 325

Figure 6 is analogous to Figure 5 with error rates (Fig. 6A) and accuracy scores (Fig. 6B) plotted as a function of feature fraction for different incipient features sizes. The random error rate sets the minimum observed error and contributes to the total error in a nonuniform way. This is because the random error term can flip values where systematic error has occurred (i.e., both sources of error can combine to produce True Positives). Figure 6C-D shows the differences in error rates and accuracy scores, respectively, between the systematic plus random error scenario shown here (Fig. 6A-B) and systematic error alone (Fig. 5A-B). The addition of random error is relatively more influential in cases where the classifier is more accurate (i.e., larger incipient features) and near endmember bedrock fractions (i.e., zero and total coverage of features). For a given incipient feature size, the minimum error added by the random error rate of 0.05 occurs at intermediate bedrock fractions and ranges from near zero for 1-m long seeds to 0.043 for 10-m long seeds. Figure 6 shows that the relative importance of random versus systematic error changes as a function of feature fraction. Because random error is the dominant term of the total error rate near the endmember cases of zero and total feature coverage, it leads to correspondingly large reductions in *nMCC* (Fig. 6D). In contrast, at intermediate bedrock fractions there is slight negative slope to *nMCC* like observed in the systematic error scenario (Fig. 5B). This is because reductions in *nMCC* induced by random error at intermediate feature fractions are: relatively smaller, approximately invariant across a broad range of fractions, and symmetrical with respect to feature fraction (Fig. 6D). While only one random error rate is shown, this example illustrates how the complex interactions between random and systematic error need to be simulated to understand their implications on pixel-level accuracy scores.

5 Discussion

Whether mapping orographic gradients in bedrock exposure (Rossi et al., 2020), characterizing precipitation controls on termite mound density (Davies et al., 2014), or inferring how wind extremes induce tree throw frequencies (Doane et al., 2023), lidar topography has revolutionized our ability to map differences in the density of fine-scale features. None of these examples used pixel-level accuracy scores in their analyses. In fact, it is not immediately apparent how well such methods would perform *even if* the authors had adopted pixel-level accuracy assessment. For those geomorphic studies that have used pixel-level accuracy scores on lidar-based classifiers (e.g., Bunn et al., 2019; Clubb et al, 2014; Milodowski et al., 2015), it is not obvious how accuracy scores are *expected to* vary as a function of feature prevalence. To help address this challenge, this paper presented a suite of synthetic scenarios that show how *F1-score* and *Matthews Correlation Coefficient* (*MCC*) perform across gradients in feature prevalence when the error structure between model and truth data are known. While the scenarios are simple, they provide insight into how well suited, and under what conditions, two of the most widely used accuracy metrics can be used when data are imbalanced (5.1). The systematic error scenarios further revealed a strong sensitivity of accuracy metrics to the shape and size of feature objects (5.2). Finally, the results from synthetic scenarios are used to provide a tentative set of best practices for using pixel-level metrics in geomorphic studies (5.3).

5.1 Accuracy assessment for imbalanced mapping tasks

One main goal of this study was to understand the sensitivity of *F1-score* and *MCC* to feature prevalence. It is useful for accuracy scores to be invariant with respect to feature fraction under a given error structure so that classified scenes can be calibrated and validated using a wide range of geomorphic settings. For example, Matthews Correlation Coefficient (*MCC*), and its normalized equivalent (*nMCC)*, readily diagnosed the case of independence between truth and model data across the full range of feature abundances (red lines in Fig. 3). In contrast, a spurious correlation between feature abundance and *F1-score* was only exacerbated by adding scene-level constraints to this case (black lines in Fig. 3). Because *F1-score* only considers True Positives, False Positive, and False Negatives, it is an asymmetric accuracy metric (Table 1). Asymmetry refers to the fact that the score is dependent on the choice of target class. All pixel-level assessments that do not consider all four components of the confusion matrix (e.g., precision, recall, F-measures, receiver operating characteristic curves) are asymmetric. Asymmetric metrics may not be problematic if one outcome is much more important than its alternative due to its consequences (e.g., a medical diagnosis). However, for many of the geomorphic mapping applications posed here, the relative importance of one class over the other is unclear (e.g., bedrock versus soil; mound versus inter-mound; incised versus un-incised). Successfully identifying both the occurrence and non-occurrence of features is important. In multi-class accuracy assessment, it is common to calculate a 'macro' *F1-score*, which is the arithmetic mean of *F1-scores* for all classes. This macro averaging can also be applied to binary tasks by calculating the *F1-score* for the alternative cases when target classes are swapped (Sokolova and Lapalme, 2009). While a macro *F1-score* for binary classification is symmetrical and easy to calculate, adoption of this approach is still relatively rare (Chicco and Jurman, 2020).

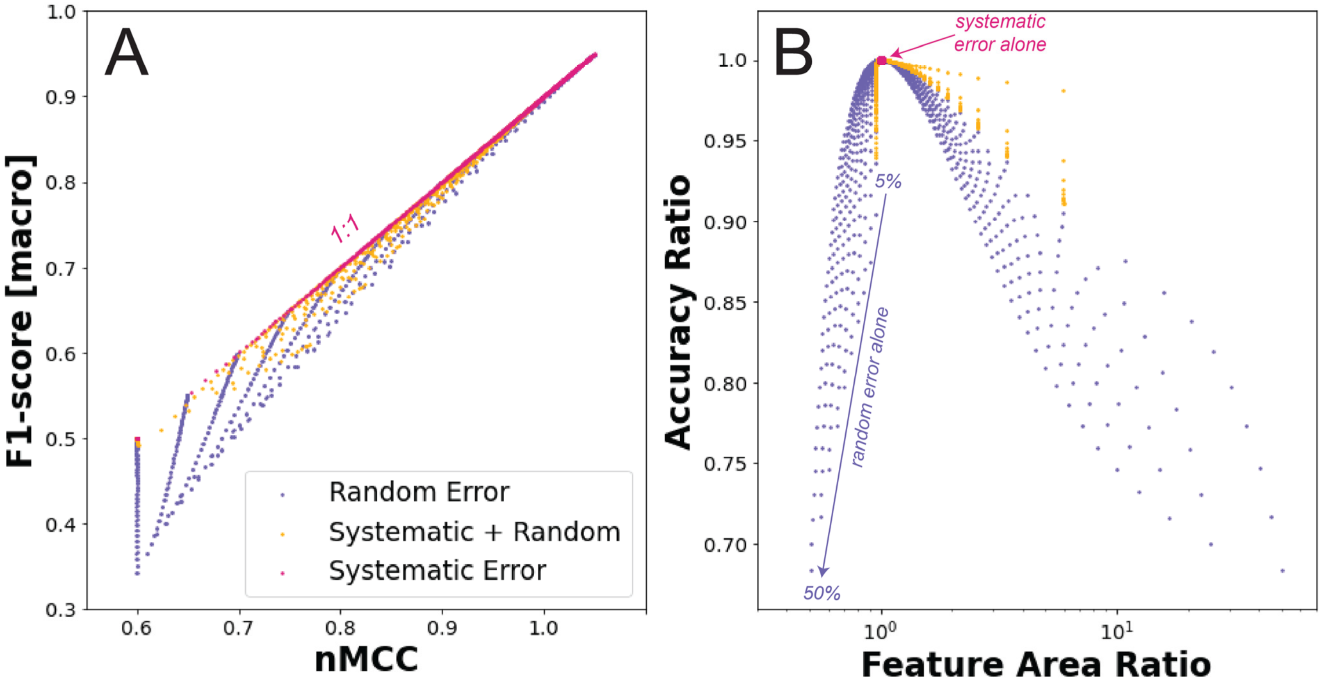

Figure 7: (A) Relationship between *nMCC* and macro *F1-score* for all the error scenarios posed in this study. (B) Ratio of accuracy scores (*nMCC* / macro *F1-score*) as a function of feature area ratios (model area / true area). In A, the macro score is the arithmetic mean of the two *F1-scores* calculated when classes are swapped. In B, the ratio of scores is plotted as a function of the ratio of feature areas to show that when the model and truth data exhibit different scene-level properties (e.g., feature areas or fraction), the macro *F1-score* produces lower values. The systematic error scenario enforced the property that model and truth data match scene-level fractions which is why they all plot at the coordinates [1,1]. The other error scenarios often produced mismatches between scene-level feature fractions. In these cases, the accuracy metrics are only equivalent when the scene-level fractions match.

Figure 7 shows how macro *F1-scores* compare to *nMCC* for each of the error scenarios considered in this paper. This modified version of *F1-score* addresses the problem of asymmetry and produces similar values to *nMCC* when the error is small. In the systematic error scenario, the scene-level fraction of bedrock in the model data is identical to the truth data. This leads to a direct correspondence between *nMCC* and macro *F1-score* (red symbols in Fig. 7). However, for the scenarios that include a fixed rate of random error, the macro *F1-scores* generally plot below the 1:1 relationship (Fig. 7A). In these scenarios, accuracy metrics are only equivalent in cases where the scene-level fractions are the same between the model and truth data (Fig. 7B). Notably, the systematic plus random error scenario produces accuracy metric ratios (Fig. 7B) closer to unity than random error alone for feature area ratios greater than one (low feature fractions). When feature area ratios are less than one (high feature fractions), accuracy ratios instead follow the trend defined by random error alone. Two important insights can be gleaned from Figure 7: (1) Even though macro *F1-score* addresses the problem of asymmetry, it penalizes random error more the *nMCC*, and (2) The mismatch between macro *F1-score* and *nMCC* is encoding disparities between scene-level and pixel-level measures of accuracy, albeit in a highly nonlinear way. Given that macro *F1-score* produces stronger sensitivity than *nMCC* to the random error scenarios (i.e., accuracy ratios < 1), *nMCC* should still be favoured as a more stable metric when calibrating and validating feature classifiers across gradients in feature prevalence. However, and despite its relative success, caution is still warranted in comparing *nMCC* across gradients in feature fraction. Uniform, random error preferentially modifies the dominant class, leading to strong reductions in accuracy near endmember cases (Fig. 4; Appendix A). Even for relatively accurate classifiers, random error limits the domain over which *nMCC* is comparable (e.g., accuracy scores for 5% random error stabilize between ~20 to 80% feature abundances; Fig. 4).

The synthetic scenarios posed in this study were motivated by tasks where differences in scene-level feature abundances are driven by differences in geomorphic setting (e.g., due to climate, ecology, material property, erosion rate). As such, the synthetic surfaces generated for this analysis assumed that feature locations were homogeneously distributed *within* each scene (like the mima mounds in Fig. 1A). The key difference *across* scenes was feature prevalence, which was used to identify how sensitive accuracy metrics are to imbalanced data. However, the sensitivity of accuracy metrics to feature fraction also provides insight into how metrics might behave when features are heterogeneously distributed within a scene (like the bedrock and gully erosion maps in Fig. 1B-C). While it is beyond the scope of this analysis to systematically explore this, a simple thought experiment using the scenes generated from this study show why within-scene heterogeneity might be important to pixel-level accuracy assessment. There are many combinations of scenes with different feature fractions that can merge into a larger one with the same feature fraction. Table 2 shows a suite of examples that each produce 50 percent feature coverage.

Table 2: Merged scenes that produce fifty percent feature area* [scene 1 percent / scene 2 percent].

	5 / 95	***10 / 90***	***15 / 85***	***20 / 80***	***25 / 75***	***30 / 70***	***35 / 65***	***40 / 60***	***45 / 55***	***50 / 50***
Random (5%)	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
	0.83 / 0.83	0.89 / 0.89	0.91 / 0.91	0.93 / 0.93	0.94 / 0.94	0.94 / 0.94	0.95 / 0.95	0.95 / 0.95	0.95 / 0.95	0.95 / 0.95
Systematic (10 m)	0.98	0.97	0.96	0.95	0.95	0.94	0.94	0.93	0.93	0.93
	0.95 / 0.86	0.95 / 0.89	0.95 / 0.90	0.94 / 0.91	0.94 / 0.91	0.94 / 0.92	0.94 / 0.92	0.94 / 0.93	0.94 / 0.93	0.93 / 0.93
Sys + Rand (10 m, 5%)	0.93	0.92	0.91	0.91	0.90	0.90	0.89	0.89	0.89	0.89
	0.80 / 0.75	0.85 / 0.80	0.87 / 0.83	0.88 / 0.85	0.89 / 0.86	0.89 / 0.87	0.89 / 0.88	0.89 / 0.88	0.89 / 0.89	0.89 / 0.89

** The top row is the nMCC of merged scenes. The bottom row is the nMCC of each individual scene that was merged.*

The merging of scenes in Table 2 helps illustrate how heterogeneous feature distributions may impact *nMCC*. For the random error scenario, the strong sensitivity to endmember cases is erased, and *nMCC* is uniform across all ten scene mixtures. For the systematic error scenario, accuracy improves for the higher feature fraction portion of the scene while accuracy marginally decreases for the lower feature fraction portions of the scene. For the systematic plus random error scenario, accuracy improves for both the higher and lower feature fraction portions of the scene. In all cases, *nMCC* is higher for the merged scenes than for their constituent components, until they converge on each other when fully homogenous. While systematic error clearly induces non-uniform mixing (i.e., merged *nMCC* varies with different constituent feature fractions), all three cases suggest that heterogeneity generally favours more stable estimates of accuracy by sampling portions of the scene with both more and less abundant features. More thorough examination of this claim is needed. Taken at face value though, these results argue that it is better to train a model on all the data at once than on individual scenes with different feature fractions, if the source of classification error is expected to be similar. However, scene-level comparisons may provide more insight into variations in the error structure of the classification model itself, which is often poorly constrained.

Taken as whole, *nMCC* should be strongly preferred over *F1-score* when building and testing classifiers across gradients in feature abundance, with heterogeneous scenes and pooling of data perhaps favouring more stable assessment. Despite this result, the two scenarios that include systematic error also suggest that asymmetry in accuracy scores is arising in response to the geometries and genesis of features. In these cases, asymmetry is not due to limitations of the accuracy metric itself, but instead a result of how features are simulated in synthetic examples. Whether the synthetic generative process (i.e., randomly distributed square features of constant size) is representative of real transitions from low to high feature fractions is an open question that likely depends on the feature of interest. Nevertheless, these synthetic examples provide an opportunity to probe how the evolution of feature geometries influence accuracy scores, a topic that is explored in much more depth below.

5.2 Size and shape of features

The focus of this paper has largely been on what to expect from pixel-level accuracy scores when a binary classifier is applied across gradients in feature prevalence. Embedded in this analysis are assumptions for how features emerge at higher abundances. Intriguingly, a negative correlation between *nMCC* and feature prevalence emerged in scenarios with systematic error, regardless of the incipient feature size (Fig. 5B; 6B). Given that *nMCC* addresses the problem of asymmetry with respect to target class (Fig. 3C-D; Fig. 4B), what causes this asymmetrical sensitivity of *nMCC* to systematic error? One likely candidate is that the simulated changes in feature prevalence entailed a corresponding change in the size and shape of feature objects. A feature object is defined here as a spatially isolated occurrence of the target class (i.e., the ones in a binary classification) enveloped by pixels of non-occurrence (i.e., the zeros in a binary classification). As features become more abundant, small objects coalesce into larger ones. This section probes the role of object size and shape on error by examining how the incipient feature shape interacts with translational error.

5.2.1 Shape and scale of incipient features

All the synthetic scenarios presented above used incipient features with square shapes and whose scale was varied using a single parameter, the incipient feature length. The square geometry was useful because squares are oriented in the same way as the regular grids being used, thus imposing a rotational symmetry to translational offsets. However, other rotationally symmetrical shapes could have been used. Figure 8 shows four alternative shapes whose rotational symmetry makes them insensitive to the direction of translational offset between truth and model data. Because these shapes are constrained by their raster representation, it is hard to create different shapes with the same area when objects are small. For the shapes ‘square’, ‘rounded’, ‘plus’, and ‘star’, all four shapes have approximately equivalent areas (< 3% difference) for shape diameters of 6, 7, 10, and 11 pixels, respectively (Fig. 8A). The number of False Positives and False Negatives to a 1-pixel offset is a function of both the object size and shape (Fig. 8B). As feature objects get larger, the relative error induced by a 1-pixel offset typically goes down. For a given object area, the relative frequency of error induced by a 1-pixel offset appears to be sensitive to the complexity of object boundaries.

To help interpret the relative trade-off between object size and shape, Figure 8C plots the *F1-scores* of the example feature objects in Figure 8A as a function of object area. Due to the symmetry of translational offset, *recall*, *precision*, and *F1-score* are equivalent for this kind of systematic error. Each of these metrics provides a measure of accuracy induced by feature shape alone, independent of the scene-level abundance of features. The error induced by a one-pixel shift between truth and model classification can be directly derived for the square case because of its simple geometry. The number of True Positives is equal to $l^2 - l$ and the number of False Positives and False Negatives are each equal to l, where l is the length of the square in integer units of pixels. Substituting these terms into equation 5 and simplifying yields an equation for *F1-score* specific to square features:

$$F1\text{-}score_{sq} = 1 - \frac{l}{l^2} \tag{8}$$

The last term in equation 8 explains why accuracy improves as a function of feature area. The area of a square increases faster than its length, thus leading to lower sensitivity to the 1-pixel offset. This ratio is equivalent to the number of pixel edges divided by the total number of pixels for a rasterized shape, which is referred to here as the edge-to-area ratio. The edge-to-area ratio can be calculated for any raster shape and sets how sensitive *F1-score* is to a translational offset. Each kind of shape differs in how the edge-to-area ratio changes as they get larger, thus defining different scaling relationships between accuracy and feature size (Fig. 8C). In general, concave shapes (i.e., 'square' and 'rounded') are more conducive to higher *F1-scores*. Concavo-convex shapes have more complex boundaries, with some shapes even showing a reduction in accuracy with increasing size (e.g., 'star' shape). Even though the synthetic scenarios used in this study assumed square seeds for their incipient features, the coalescing of these incipient shapes into larger objects means that complex boundaries, and thus increasing edge-to-area ratios emerge as feature prevalence increases.

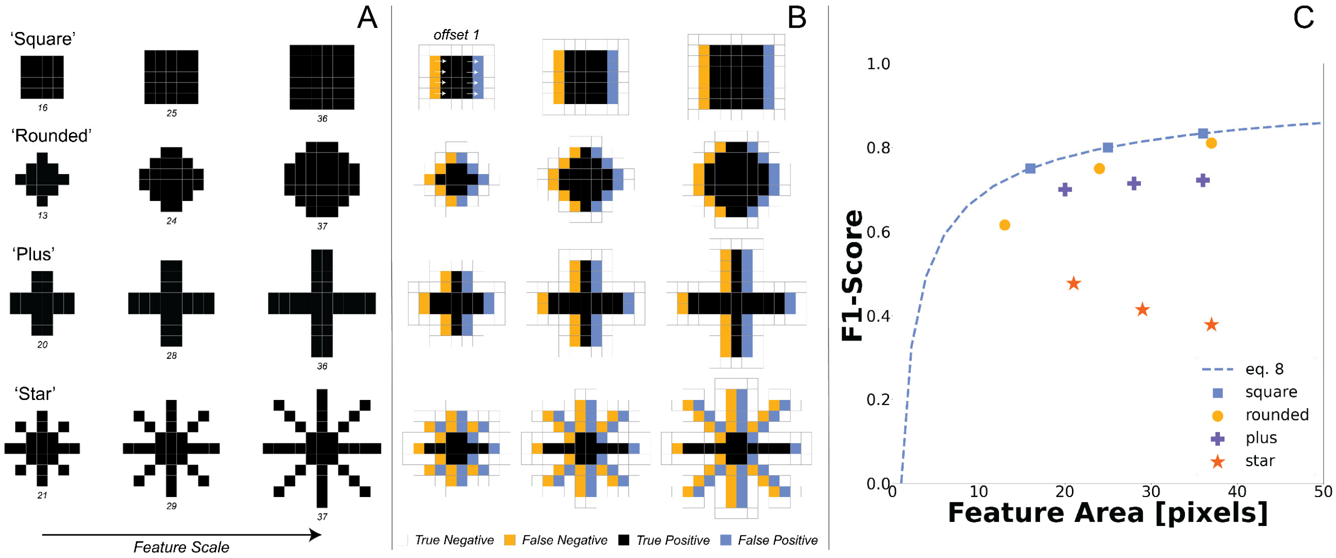

Figure 8: (A) The shape and scale of incipient feature objects directly affects (B) the subsequent frequencies of False Negatives (yellow) and False Positives (blue) to a 1-pixel, translational offset in model classification, (C) which also results in different scaling relationships between object areas and *F1-score*. In A, four different objects are shown that have either convex (i.e., square, rounded) or concavo-convex (i.e., plus, star) boundaries with respect to the matrix. The object area is reported below each shape in pixels. Note that the smallest 'rounded' example is not actually round, but a rotated square. In B, error classes are shown for a 1-pixel shift to the right. Because shapes are all rotationally symmetric with respect to the four cardinal directions, error rates do not depend on the direction of the shift. Only true negatives that share an edge with the other classes are shown. In C, the *F1-score* for each of the sixteen shapes are plotted as a function of the object area. The function describing how object area and *F1-score* varies for square features (eq. 8) is also plotted as a dashed line for reference.

5.2.2 Shape and scale of emergent features

In the synthetic scenarios presented above, the minimum feature size is set by the incipient feature length (i.e., 1 to 10 pixels). Because incipient features are placed on the surface randomly, more complex objects are produced where incipient features overlap by chance. To illustrate the implications of this, Figure 9A shows examples of individual objects that can be generated using square seeds. Examples are organized by incipient feature size (rows) and object areas (columns). Adjacent to each object is the error induced by a 1-pixel shift to the right, with its corresponding *F1-score* reported above it. Note that individual objects are not necessarily rotationally symmetric. If an object has a preferred orientation, then error will be enhanced for objects where the long axis is parallel to the translational offset and reduced for objects where the long axis is perpendicular to the translational offset. In practice, the sensitivity of error to object orientation is not realized in the synthetic scenarios above because the random placement of features results in objects without a preferred orientation.

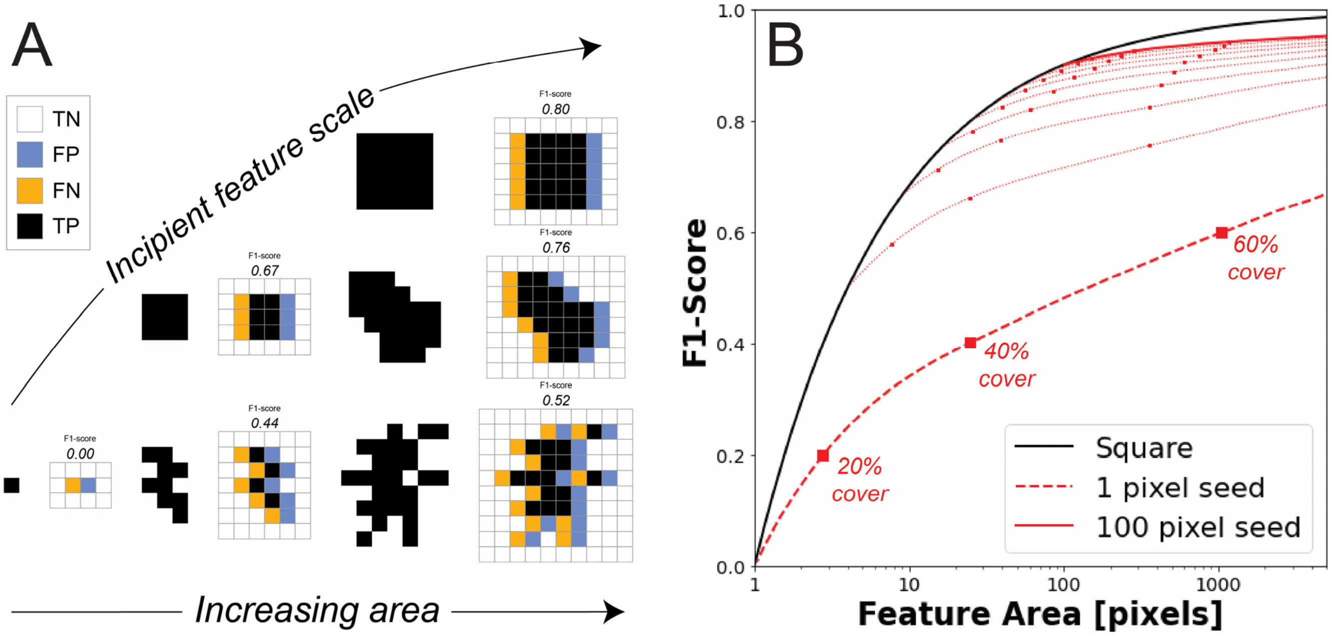

Figure 9: (A) For a given object area, the frequency of False Positives and False Negatives differs among incipient objects and the emergent objects that coalesce from smaller ones, such that (B) *F1-scores* increase with average object area more slowly than square objects do in response to a 1-pixel offset. In A, six permissible object shapes are shown for three different incipient feature sizes (rows) and three different object areas (columns). The incipient feature shape both controls the minimum object size and the complexity of object boundaries. Smaller incipient features can produce more complex shapes and higher error rates for a given feature size (see associated *F1-scores*). In B, the *F1-score* is plotted as a function of average object area for the systematic error scenario (Fig. 5). Markers show values at three different feature fractions. The black line is the function describing how *F1-score* responds to a 1-pixel offset to an individual square object (eq. 8).

While the examples shown in Figure 9A reiterate the point that error is reduced for larger objects with simpler shapes in response to a 1-pixel offset, it still does not show how object properties are varying in the synthetic scenarios presented above. Figure 9B plots the *F1-score* as a function of the mean object area for the systematic error scenario. To calculate object areas, the binary map of features (i.e., pixel values equal to one) is segmented into objects. Object segmentation is based on adjacency of the target feature class in at least one of its eight neighbours (see examples in Fig. 9A). Objects can contain holes, but these holes do not contribute to their object area. After segmenting the scene into objects, the average object area is calculated and linked to the *F1-scores* reported earlier (Fig. 5). Figure 9B shows that *F1-score* generally improves with increasing object area, albeit in a way that is strongly mediated by the incipient feature size. All lines intersect with the function describing *F1-*

score for square features (eq. 8; solid black line) for the limiting case where there is only one object in the scene. For any given incipient feature size though, *F1-score* quickly drops off this function due to the increasing complexity of object boundaries. There is a monotonic increase in *F1-score* with average object area and feature prevalence (markers in Fig. 9B) regardless of the incipient feature size. The scenarios above are not producing shapes like the 'stars' shown in Figure 8. Larger features do lead to higher *F1-scores* (Fig. 9B). It was already shown that placing larger features in the landscape improves accuracy in response translational offset (Figs. 5-6). As such, the negative trends in *nMCC* shown in Figures 5B and 6B suggest that the net result of increasing the size of features, for a given incipient seed, is outweighed by the complexity of feature boundaries generated by coalescing them. This analysis suggests that it is paramount to understand the scaling properties of features as they become more prevalent to understand how accuracy scores may be affected by small co-registration errors. Finally, the sensitivity of accuracy metrics to the size and shape of individual features begs important questions as to how stable accuracy metrics are to increasing spatial resolution. As airborne remote sensing is supplemented and superseded by drone-based mapping, there is good reason to believe that the sizes and shapes of better resolved features may change, and thus influence how binary classifiers perform.

5.3 Recommendations and future directions

Many geomorphic tasks share the need for binary classifiers that perform well across gradients in feature abundance. Whether constraining the density of landslide scars, channel erosion, bedrock outcrops, or pit-mound features, geomorphic studies often rely on fine-scale mapping to determine how feature size, extent, and prevalence respond to differences in environmental forcing. There is a general need for classifiers that successfully handle imbalanced data. This paper set out to understand how two widely used pixel-level accuracy metrics perform across gradients in feature prevalence. By using synthetic examples where the error structure of the data is known, heuristics can be developed for best practices when the research design specifically calibrates and validates binary classifiers across gradients in feature abundance. Four key recommendations emerged:

(1) *Matthews Correlation Coefficient*, and its normalized equivalent (*nMCC*), are much better suited than *F1-score* to comparing accuracy scores when feature abundances vary across classified scenes. Even after addressing the problem of asymmetry, macro *F1-score* tends to over-penalize random error.

(2) For random error, caution is warranted in interpreting *nMCC* near the endmember cases of zero and full feature coverage because random error preferentially modifies the dominant class. Though scores are relatively invariant only between ~20-80% feature coverage, this domain might be expanded for scenes with more heterogeneous feature distributions.

(3) For systematic error, *nMCC* is strongly sensitive to the size and shape of individual objects. Larger objects with simpler boundaries are less sensitive to this kind of error because their edge-to-area ratios are small. As such, it is important to characterize both co-registration uncertainty and the attributes of the individual objects being mapped.

(4) Before training and testing classifiers on imbalanced data, it is essential to establish baseline expectations for how pixel-level accuracy scores respond to potential sources of error over the range of feature abundances observed. This can be accomplished through numerical simulation.

Simulating a suite of simple scenarios with a known error structure and uniform incipient seeds provided some insight into how pixel-level accuracy metrics behave across gradients in feature prevalence. Real-world applications are decidedly more complex. In the scenarios presented here, increased feature density was simulated by randomly distributing the nuclei of incipient features within the model domain. Such a treatment may be relevant to some applications but is clearly limited. Figure 1 anticipated three clear limitations of simulating features in this way. Many features show evidence for: a characteristic size and spacing (e.g., mima mounds in Fig. 1A), size distributions spread across a wide range of scales (e.g., bedrock exposure in Fig. 1B), and anisotropy (e.g., gully erosion in Fig. 1C). As such, more work is needed to understand how pixel-level accuracy metrics perform on imbalanced data that exhibit these properties. To this end, three promising future research directions are:

(1) As **landscape evolution modelling** attempts to keep pace with increasingly higher resolution observations (Tucker & Hancock, 2010), it also has wide potential for error analysis. Instead of randomly generating features, numerical models can produce more realistic feature distributions that are derived from the relevant geomorphic transport laws at play (Dietrich et al., 2003). A process-based approach towards error assessment could be used to identify under what conditions binary classifiers can be reliably compared across gradients in feature fraction.

(2) Pixel-level accuracy scores are built on the confusion matrix, which does not retain the **spatial autocorrelation** structure or the **semantic content** of feature objects. Given the importance of the size and shape of features to some error scenarios, the path forward may lie in multi-scale, object-based image analysis (e.g., Drăguţ and Eisank, 2011). Object-based image analysis is on the cutting edge of feature extraction from remote sensing data (Hossain and Chen, 2019). Yet, how to reliably evaluate the accuracy of image segmentation algorithms will require creative re-thinking and re-tooling of standard pixel-level accuracy scores (Cai et al., 2018).

(3) Both opportunities above emphasize the over-arching challenge of the rapidly changing landscape of increasing **spatial resolution** data. Higher resolution data both impacts the practical challenge of co-registration error as well as highlights the more theoretical challenge of semantic vagueness, or the notion that feature boundaries may not be

sharply defined (Sofia, 2020). As data resolution increases, traditional methods in image segmentation and binary classification may require new approaches (Zheng and Chen, 2023).

On the one hand, this paper is a call to action on adopting standard methods from the data sciences into surface processes research. On the other hand, geomorphic questions provide a diversity of real-world use-cases where these ‘standard’ methods can be put to the test and new methods can be developed. As machine learning approaches towards geomorphic mapping proliferate, better understanding is needed on how these methods will perform on the scientific tasks that are currently driving surface processes research forward.

6 Conclusions

Pixel-level accuracy assessment provides a powerful tool for understanding how well classifiers built from high resolution topography are performing. To be most useful, the limitations of commonly used metrics like *precision*, *recall*, and *F1-score* need to be considered. Classification tasks that span large gradients in feature abundance are particularly vulnerable to biases in these metrics because data is imbalanced and the choice of target class matters. More robust metrics like *MCC* and *nMCC* largely address these methodological challenges. However, caution is still warranted in comparing pixel-level scores across gradients in feature density and extent. If error is random and uniform across scenes, then *nMCC* will dramatically worsen near endmember cases because the more prevalent class is preferentially modified, though this effect may be mediated by pooling data from many different scenes. If the model is systematically offset from the truth grid, then an asymmetrical sensitivity of *nMCC* can arise depending on assumptions for the genesis and growth of individual features. As the size of individual features increases, there is lower sensitivity to systematic offset. However, if the shapes of features are also getting more complex, then the increased edge to area ratio of individual features can counteract and exceed improvements in accuracy associated with larger feature sizes. Though pixel-level metrics used in the machine learning and remote sensing community should be more widely adopted in geomorphic research, further work is needed to understand how different sources of error might decouple pixel-level from scene-level measures of accuracy.

Appendix A: Random error and accuracy metrics

Section 4.1 reported how pixel-level accuracy scores vary as a function of bedrock fraction for a fixed rate of random error. While the synthetic surfaces were generated using Python, the results shown in Figure 4 can be directly derived from the mean random error rate ($\bar{e}_r$) and true feature fraction (f_f) analytically. Under this scenario, the probability of flipping either class is independent of the prevalence and location of features such we can define the average frequencies for all four components of the confusion matrix. The relative frequencies of each outcome are the product of the average rate of error (or non-error) and the average abundance of the true class. For example, the True Positives reflect both the probability of the feature occurring

(f_f) and the probability of not being flipped in the model due to random error (i.e., $1-\bar{e}_r$). The frequencies of all four classification outcomes are:

$$f_{TP} = (1-\bar{e}_r)f_f \tag{A1}$$

$$f_{FP} = \bar{e}_r f_f \tag{A2}$$

$$f_{FN} = \bar{e}_r\left(1-f_f\right) \tag{A3}$$

$$f_{TN} = (1-\bar{e}_r)\left(1-f_f\right) \tag{A4}$$

Because we also know that the feature fraction in the model (f_{fm}) must equal the sum of the fractions of True Positives and False Negatives, these equations yield the relationship:

$$f_{fm} = (1-\bar{e}_r)f_f + \bar{e}_r\left(1-f_f\right) \tag{A5}$$

Equation A5 can be rearranged and simplified to describe how the model feature fraction is related to the true feature fraction:

$$f_{fm} = (1-2\bar{e}_r)f_f + \bar{e}_r \tag{A6}$$

The relationships shown in Figure 4A (main text) are equivalent to equation A6 for different error rates. That the Python-generated scenes match the analytical solution indicates that the domain used for these synthetic scenes is large enough to adequately sample population statistics. Note that equation A6 provides a prediction for the relationship between true and model bedrock fractions only if error is uniform and random across scenes. In such cases, the average error rate can be directly inferred from both the slope and y-intercept of the regression.

Because pixel-level accuracy scores can be derived directly from the confusion matrix, the simplified assumptions of random, uniform error also facilitate prediction for how *F1-score* and *nMCC* will vary with the true feature fraction. Substituting the values from eqs. A1-A4 into equation 5 (main text) yields:

$$F1\text{-}score = \frac{2f_f(1-\bar{e}_r)}{2f_f(1-\bar{e}_r)+\bar{e}_r} \tag{A7}$$

which is equivalent to the numerically generated black curves in Figure 4B. Similarly, substituting eqs. A1-A4 into equation 6 (main text) yields:

$$MCC = \frac{\sqrt{f_f}\times\sqrt{1-f_f}\times(1-2\bar{e}_r)}{\sqrt{f_f+\bar{e}_r-\bar{e}_r{}^2-f_f{}^2-4\bar{e}_r f_f-4\bar{e}_r f_f{}^2-4\bar{e}_r{}^2 f_f-4\bar{e}_r{}^2 f_f{}^2}} \tag{A8}$$

which is equivalent to the numerically generated red curves in Figure 4B. Though the expression for *MCC* under random, uniform error is complex, it reveals why there is strong and symmetrical sensitivity near the endmember cases of zero and all bedrock. The numerator in eq. A8 decreases faster than the denominator near endmember cases regardless of the average error rate. Since f_f and $1-f_f$ are complementary and $\bar{e}_r$ is assumed to be constant, this reduction in *MCC* is also symmetrical around an optimal bedrock fraction of 0.5.

Data Availability

Figure 1 elevation data was downloaded from OpenTopography (2010 Channel Islands Lidar Collection, 2012; Anderson et al., 2012; Reed, 2006). Figure 2 and Table 1 are based on the bedrock mapping at site P01 from Rossi et al. (2020). Maps for 1-m truth and model data at this site can be accessed at https://github.com/mwrossi/cfr_extremes. These classified maps are based on 2018 Pictometry® orthomosaicked air photos purchased by Boulder County and airborne lidar data acquired by the National Center for Airborne Laser Mapping for the Boulder Creek Critical Zone Observatory (Anderson et al., 2012). Synthetic surfaces presented in Figures 3-9 were built in Python. Scripts can be accessed at https://github.com/mwrossi/feature-mapping-accuracy. Once through review, the main code will continue to be hosted on Github, but scripts and files used for generating each figure will be archived on Figshare.

Competing interests

The author declares that there is no conflict of interest.

Acknowledgements

This project was supported by funding from the Geomorphology and Land-Use Dynamics Program at the National Science Foundation (EAR-1822062) and benefited from discussions with Bob Anderson, Suzanne Anderson, Roman DiBiase, Brittany Selander, and Greg Tucker. The author thanks one anonymous reviewer and Stuart Grieve for their thoughtful and constructive reviews.

References

2010 Channel Islands Lidar Collection, United States Geological Survey, distributed by OpenTopography [data set], https://doi.org/10.5069/G95D8PS7, 2012.

Ågren, A. M., Larson, J., Paul, S. S., Laudon, H., and Lidberg, W.: Use of multiple LIDAR-derived digital terrain indices and machine learning for high-resolution national-scale soil moisture mapping of the Swedish forest landscape, Geoderma, 404, 115280, https://doi.org/10.1016/J.GEODERMA.2021.115280, 2021.

Anderson, S.P., Qinghua, G., and Parrish, E.G.: Snow-on and snow-off Lidar point cloud data and digital elevation models for study of topography, snow, ecosystems and environmental change at Boulder Creek Critical Zone Observatory, Colorado, National Center for Airborne Laser Mapping, distributed by OpenTopography [data set], https://doi.org/10.5069/G93R0QR0, 2012.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., and Nielsen, H.: Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics, 16, 412–424, https://doi.org/10.1093/BIOINFORMATICS/16.5.412, 2000.

Barnhart, K. R., Tucker, G. E., Doty, S. G., Glade, R. C., Shobe, C. M., Rossi, M.W., and Hill, M. C.: Projections of landscape evolution on a 10,000 year timescale with assessment and partitioning of uncertainty sources, J. Geophys. Res.-Earth, 125, e2020JF005795, https://doi.org/10.1029/2020JF005795, 2020.

Bertin, S., Jaud, M., and Delacourt, C.: Assessing DEM quality and minimizing registration error in repeated geomorphic surveys with multi-temporal ground truths of invariant features: Application to a long-term dataset of beach topography and nearshore bathymetry, Earth Surf. Process. Landforms, 47, 2950-2971, https://doi.org/10.1002/ESP.5436, 2022.

Bunn, M. D., Leshchinsky, B. A., Olsen, M. J., and Booth, A.: A simplified, object-based framework for efficient landslide inventorying using LIDAR digital elevation model derivatives, Remote Sens., 11, 303, https://doi.org/10.3390/rs11030303, 2019.

Cai, L., Shi, W., Miao, Z., and Hao, M.: Accuracy assessment measures for object extraction from remote sensing images, Remote Sens., 10, 303, https://doi.org/10.3390/rs10020303, 2018

Chicco, D. and Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics, 21, 1-13, https://doi.org/10.1186/S12864-019-6413-7, 2020.

Chicco, D., Warrens, M. J., and Jurman, G.: The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment, IEEE Access, 9, 78368–78381, https://doi.org/10.1109/ACCESS.2021.3084050, 2021a.

Chicco, D., Tötsch, N., and Jurman, G.: The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation, BioData Min., 14, 1–22, https://doi.org/10.1186/S13040-021-00244-Z, 2021b.

Chinchor, N.: MUC-4 evaluation metrics, in: Proceedings of MUC-4 - the 4th Conference on Message Understanding, McLean, VA, 16-18 June 1992, 22-29, https://doi.org/10.3115/1072064.1072067, 1992.

Clubb, F. J., Mudd, S. M., Milodowski, D. T., Hurst, M. D., and Slater, L. J.: Objective extraction of channel heads from high-resolution topographic data, Water Resour. Res., 50, 4283–4304, https://doi.org/10.1002/2013WR015167, 2014.

Cunningham, D., Grebby, S., Tansey, K., Gosar, A., and Kastelic, V.: Application of airborne LiDAR to mapping seismogenic faults in forested mountainous terrain, southeastern Alps, Slovenia, Geophys. Res. Lett., 33, https://doi.org/10.1029/2006GL027014, 2006.

Davies, A. B., Levick, S. R., Asner, G. P., Robertson, M. P., Van Rensburg, B. J., Parr, C. L., Davies, A. B., Robertson, M. P., and Van Rensburg, B. J.: Spatial variability and abiotic determinants of termite mounds throughout a savanna catchment, Ecography, 37, 852–862, https://doi.org/10.1111/ecog.00532, 2014.

DiBiase, R. A., Heimsath, A. M., and Whipple, K. X.: Hillslope response to tectonic forcing in threshold landscapes, Earth Surf. Process. Landforms, 37, 855–865, https://doi.org/10.1002/esp.3205, 2012.

Dietrich, W. E., Bellugi, D. G., Sklar, L. S., Stock, J. D., Heimsath, A. M., and Roering, J. J.: Geomorphic Transport Laws for Predicting Landscape form and Dynamics, Geophys. Monogr. Ser., 135, 103–132, https://doi.org/10.1029/135GM09, 2003.

Doane, T. H., Yanites, B. J., Edmonds, D. A., and Novick, K. A.: Hillslope roughness reveals forest sensitivity to extreme winds, Proc. Natl. Acad. Sci. U. S. A., 120, e2212105120, https://doi.org/10.1073/PNAS.2212105120, 2023.

Drăguţ, L. and Eisank, C.: Object representations at multiple scales from digital elevation models, Geomorphology, 129, 183-189, https://doi.org/10.1016/j.geomorph.2011.03.003, 2011.

Hossain, M. D. and Chen, D.: Segmentation for Object-Based Image Analysis (OBIA): A review of algorithms and challenges from remote sensing perspective, ISPRS J. Photogramm., 150, 115-134, https://doi.org/10.1016/j.isprsjprs.2019.02.009, 2019.

Jaboyedoff, M., Oppikofer, T., Abellán, A., Derron, M. H., Loye, A., Metzger, R., and Pedrazzini, A.: Use of LIDAR in landslide investigations: a review, Nat. Hazards, 61, 5–28, https://doi.org/10.1007/S11069-010-9634-2, 2012.

Korzeniowska, K., Pfeifer, N., and Landtwing, S.: Mapping gullies, dunes, lava fields, and landslides via surface roughness, Geomorphology, 301, 53-67, https://doi.org/10.1016/j.geomorph.2017.10.011, 2018.

Levick, S. R., Asner, G. P., Chadwick, O. A., Khomo, L. M., Rogers, K. H., Hartshorn, A. S., Kennedy-Bowdoin, T., and Knapp, D. E.: Regional insight into savanna hydrogeomorphology from termite mounds, Nat. Commun., 1, 65, https://doi.org/10.1038/ncomms1066, 2010.

Matthews, B. W.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta - Protein Struct., 405, 442–451, https://doi.org/10.1016/0005-2795(75)90109-9, 1975.

Milodowski, D. T., Mudd, S. M., and Mitchard, E. T. A.: Topographic roughness as a signature of the emergence of bedrock in eroding landscapes, Earth Surf. Dyn., 3, 483–499, https://doi.org/10.5194/ESURF-3-483-2015, 2015.

Morell, K. D., Regalla, C., Leonard, L. J., Amos, C., and Levson, V.: Quaternary rupture of a crustal fault beneath Victoria, British Columbia, Canada, GSA Today, 27, 4–10, https://doi.org/10.1130/GSATG291A.1, 2015.

Passalacqua, P., Belmont, P., Staley, D. M., Simley, J. D., Arrowsmith, J. R., Bode, C. A., Crosby, C., DeLong, S. B., Glenn, N. F., Kelly, S. A., Lague, D., Sangireddy, H., Schaffrath, K., Tarboton, D. G., Wasklewicz, T., and Wheaton, J. M.: Analyzing high resolution topography for advancing the understanding of mass and energy transfer through landscapes: A review, Earth-Science Rev., 148, 174–193, https://doi.org/10.1016/J.EARSCIREV.2015.05.012, 2015.

Pavlis, T. L. and Bruhn, R. L.: Application of LIDAR to resolving bedrock structure in areas of poor exposure: An example from the STEEP study area, southern Alaska, GSA Bull., 123, 206–217, https://doi.org/10.1130/B30132.1, 2011.

Pirotti, F. and Tarolli, P.: Suitability of LiDAR point density and derived landform curvature maps for channel network extraction, Hydrol. Process., 24, 1187–1197, https://doi.org/10.1002/HYP.7582, 2010.

Prakash, N., Manconi, A., and Loew, S.: Mapping Landslides on EO Data: Performance of deep learning models vs. traditional machine learning models, Remote Sens., 12, 346, https://doi.org/10.3390/RS12030346, 2020.

Reed, S.: Merced, CA: Origin and evolution of the Mima mounds, National Center for Airborne Laser Mapping, distributed by OpenTopography [data set], https://doi.org/10.5069/G93B5X3Q, 2006.

Reed, S. and Amundson, R.: Using LIDAR to model Mima mound evolution and regional energy balances in the Great Central Valley, California, Spec. Pap. Geol. Soc. Am., 490, 21–41, https://doi.org/10.1130/2012.2490(01), 2012.

van Rijsbergen, C. J.: Foundation of evaluation. J. Doc., 30, 365-373, https://doi.org/10.1108/eb026584, 1974.

Roering, J. J., Marshall, J., Booth, A. M., Mort, M., and Jin, Q.: Evidence for biotic controls on topography and soil production, Earth Planet. Sci. Lett., 298, 183–190, https://doi.org/10.1016/J.EPSL.2010.07.040, 2010.

Roering, J. J., Mackey, B. H., Marshall, J. A., Sweeney, K. E., Deligne, N. I., Booth, A. M., Handwerger, A. L., and Cerovski-Darriau, C.: “You are HERE”: Connecting the dots with airborne lidar for geomorphic fieldwork, 200, 172–183, https://doi.org/10.1016/j.geomorph.2013.04.009, 2013.

Rossi, M. W., Anderson, R. S., Anderson, S. P., and Tucker, G. E.: Orographic Controls on Subdaily Rainfall Statistics and Flood Frequency in the Colorado Front Range, USA, Geophys. Res. Lett., 47, e2019GL085086, https://doi.org/10.1029/2019GL085086, 2020.

Sofia, G.: Combining geomorphometry, feature extraction techniques and Earth-surface processes research: The way forward, 355, 107055, https://doi.org/10.1016/J.GEOMORPH.2020.107055, 2020.

Sokolova, M. and Lapalme, G.: A systematic analysis of performance measures for classification tasks, Inf. Process. Manag., 45, 427-437, https://doi.org/10.1016/j.ipm.2009.03.002, 2009.

Tucker, G. E. and Hancock, G. R.: Modelling landscape evolution, Earth Surf. Process. Landforms, 35, 28–50, https://doi.org/10.1002/ESP.1952, 2010.

Wang, Y., Fang, Z., and Hong, H.: Comparison of convolutional neural networks for landslide susceptibility mapping in Yanshan County, China, Sci. Total Environ., 666, 975–993, https://doi.org/10.1016/J.SCITOTENV.2019.02.263, 2019.

Zheng, X. and Chen, T.: High spatial resolution remote sensing image segmentation based on the multiclassification model and the binary classification model, Neural Comput. Appl., 35, 3597–3604, https://doi.org/10.1007/S00521-020-05561-8, 2023.