Geomorphology 381 (2021) 107660

Contents lists available at ScienceDirect

Geomorphology

journal homepage: www.elsevier.com/locate/geomorph

Comparing methods of landslide data acquisition and susceptibility modelling: Examples from New Zealand

Hugh G. Smith *, Raphael Spiekermann, Harley Betts, Andrew J. Neverman

Manaaki Whenua – Landcare Research, Palmerston North, New Zealand

ARTICLE INFO

Article history:
Received 17 September 2020
Received in revised form 6 January 2021
Accepted 14 February 2021
Available online 19 February 2021

Keywords:
Landslide susceptibility
Shallow landslides
Landslide inventory
Rainfall-induced landslides
Object-based image analysis (OBIA)
Semi-automated mapping

ABSTRACT

The acquisition of landslide inventory data remains an important challenge for landslide susceptibility modelling. For rainfall-induced landslides, comprehensive mapping may be hindered by the size of storm-affected areas and large number of landslides generated as well as the time and costs involved in preparing multi-temporal inventories. In New Zealand, storm events trigger hundreds to thousands of shallow landslides, causing significant damage to land and infrastructure as well as impacts on freshwater and marine environments. Despite this, there are few quantitative assessments of shallow landslide susceptibility to inform targeting of control measures. Here, we compare the effect of using landslide inventories assembled from a) manual versus semi-automated mapping and b) event versus multi-temporal records on the performance of two widely applied methods for landslide susceptibility modelling, namely logistic regression and random forest classification. Evaluation of object-based image analysis (OBIA) for semi-automated landslide mapping showed mixed results, where producer's and user's accuracies ranged 62–81 and 45–55%, respectively, without manual refinement. However, the relative reduction of 6–11% in susceptibility model predictive performance based on area under receiver operating characteristic curves (AUC) using OBIA (AUC = 0.63–0.75) versus manual (AUC = 0.67–0.81) inventories with different variable combinations was low in comparison, and the spatial patterns in modelled susceptibility were generally similar. The random forest model produced slightly better prediction performance compared with logistic regression based on cross-validation within the same study area. However, this was reversed and logistic regression mostly outperformed random forest when the models were fitted and tested with data from different study areas. Model predictive performance for event versus multi-temporal records was comparable. Our results highlight both the challenges associated with semi-automated landslide detection over large areas as well as the opportunity to use OBIA for efficient data collection without necessarily compromising the resulting susceptibility maps. This potentially overcomes significant time and cost impediments to the preparation of landslide inventories that continue to hinder quantitative landslide susceptibility assessment.

1. Introduction

Most landslides in New Zealand are rapid, shallow slides and flows that occur in soil and regolith in response to storm rainfall (Crozier, 1996; Basher, 2013). Hundreds to thousands of these shallow landslides may be triggered by infrequent large storms, so called multiple-occurrence regional landslide events (MORLEs) (Crozier, 2017). These events have well-documented impacts in terms of erosion damage to pastoral land (Dymond et al., 2006) and associated agricultural production losses (Krausse et al., 2001; Dominati et al., 2014). Shallow landslides are considered the dominant erosion process in New Zealand's hill country terrain, defined as land less than 1000 m in elevation with slopes typically between 20 and 30° on a variety of rock types, but most commonly in areas dominated by weak sedimentary rocks from the Tertiary and Quaternary (Crozier, 2010). Moreover, shallow landslide erosion, accelerated by land clearing, typically makes the largest erosion process contribution to sediment budgets for hill country catchments over multi-decadal timescales (Trustrum et al., 1999; Dymond et al., 2016). Despite the widespread occurrence of shallow landslides, there have been few quantitative assessments of rainfall-induced shallow landslide susceptibility in New Zealand (e.g. Dymond et al., 2006; Schicker and Moon, 2012). This relative paucity of landslide susceptibility information has been noted internationally where there is a distinct geographical bias towards Asia and Europe with the least number of studies in Oceania (Reichenbach et al., 2018).

The number of landslides and size of areas affected by MORLEs is a significant challenge in the acquisition of landslide data for susceptibility assessments. This often prevents comprehensive mapping of storm-impacted areas, restricting the development of landslide inventories due to the time and costs involved. In New Zealand, individual shallow landslide source areas (scars) in hill country are typically small (approximately 50–100 m^2 in median scar size). As a result, we require very high resolution (VHR) imagery to enable 1) accurate detection of

* Corresponding author.
E-mail address: smithh@landcareresearch.co.nz (H.G. Smith).

https://doi.org/10.1016/j.geomorph.2021.107660
0169-555X/

individual landslide features and 2) separation of landside scar and debris deposits for use in landslide susceptibility modelling. Semi-automated mapping techniques offer a potential means of efficiently acquiring such landslide data (Guzzetti et al., 2012). To date, several studies have demonstrated the potential for object-based image analysis (OBIA) to detect landslide features (Martha et al., 2010; Blaschke et al., 2014; Kurtz et al., 2014; Hölbling et al., 2015, 2016).

Multi-temporal landslide inventories are considered the optimal source of landslide information for statistical landslide susceptibility modelling (Reichenbach et al., 2018). By contrast, these authors suggest caution in using individual storm inventories due to their dependency on the pattern and extent of rainfall triggering events. Nonetheless, event inventories remain the second most frequently used inventory type for susceptibility modelling (Reichenbach et al., 2018). Multi-temporal inventories also present challenges, not least of which are the time and cost associated with their preparation (Guzzetti et al., 2012). In New Zealand, the detectability of small and shallow landslides can reduce rapidly with time as landslides re-vegetate and scar margins degrade, particularly when mapping is reliant on historic aerial photography that can be low resolution (De Rose, 2013) or from up to a decade following the triggering event (Betts et al., 2017). This may result in under-estimation of the full extent of shallow landslide occurrence. In contrast, acquiring landslide inventories soon after storm events using VHR imagery increases the likelihood that all landslides within a study area will be included.

The few published studies of landslide susceptibility in New Zealand employed heuristic and statistically based classification approaches. Dymond et al. (2006) developed a heuristic shallow landslide susceptibility model where a pixel was defined as susceptible if it a) exceeds a pre-defined slope threshold that varied according to rock type (range 24–28°) and b) does not have closed canopy woody vegetation cover. Schicker and Moon (2012) employed logistic regression and weights of evidence methods using landslide data for the Waikato region on the North Island from geological mapping and the GeoNet landslide database (Rosser et al., 2017). However, the authors note that these data are biased towards large landslides and the inventory comprised only 229 landslides for a 25,000 km^2 region. Kritikos and Davies (2015) examined rainfall-triggered shallow landslides in the mountain terrain of the western Southern Alps of New Zealand and developed a susceptibility model with 706 landslides using fuzzy logic that adopted both data and knowledge-driven approaches.

Here, we compare predictions of shallow landslide susceptibility using two statistical susceptibility models with landslide inventories derived from different mapping techniques and temporal intervals. OBIA methods have only recently been used to derive landslide inventories for susceptibility modelling (Amato et al., 2019). These authors applied OBIA mapping in combination with manual refinement of the resulting inventory for use in modelling. In contrast, our study is the first to compare susceptibility model performance using landslide inventories derived from manual versus OBIA mapping techniques. We do not manually refine our OBIA results and instead use manual scar area mapping to independently evaluate OBIA classification performance. This ensures that OBIA results directly reflect OBIA classification performance rather than a combination of OBIA and manual procedures that depends on the extent of manual editing. This approach allows us to examine the balance between rapid, high-volume data acquisition made possible by semi-automated methods versus potential reductions in mapping accuracy and the effect on landslide susceptibility assessment.

Numerous landslide susceptibility studies report comparisons of different statistical methods (e.g. Rossi et al., 2010; Pradhan, 2013; Goetz et al., 2015; Chen et al., 2017; Pourghasemi and Rahmati, 2018). In their review, Reichenbach et al. (2018) noted this tendency in the literature for excessive experimentation with statistical methods that they considered largely unjustified. For this reason, we do not seek to compare a range of statistical models here. Instead, we select two models, namely logistic regression and random forest, on the basis that logistic regression is the most widely applied model in the landslide susceptibility literature (Reichenbach et al., 2018) and thus maximises comparability, while machine learning methods such as random forest have received increasing attention and been found to perform well compared with other methods (Goetz et al., 2015; Youssef et al., 2016; Pourghasemi and Rahmati, 2018).

The present study has the following objectives: 1) evaluate the accuracy of OBIA for semi-automated landslide mapping following storm events; 2) compare susceptibility model performance based on landslide inventories produced by manual versus OBIA mapping techniques; and 3) assess susceptibility models based on event versus multi-temporal landslide inventories to better understand how inventory type affects susceptibility model performance. We draw on shallow landslide inventories for six study areas located in hill country environments on the North Island of New Zealand. These comprise three storm event and three multi-temporal inventories derived from VHR satellite imagery and aerial photography. Compared with the few previous statistical studies of landslide susceptibility in New Zealand, we draw on a much larger dataset of over 53,000 shallow landslides.

2. Methods

2.1. Study areas

The shallow landslide mapping and susceptibility modelling focused on six study areas ranging in size from 9.4 to 178 km^2 (Table 1). These areas are located across the central North Island of New Zealand which contains extensive hill country terrain (Fig. 1). All study areas

Table 1
Summary information for the study areas contributing shallow landslide data for susceptibility modelling.

Study area no.	Location	Study area (km^2)	Period	Rainfall (mm) & [duration, hours]	ARI[a] (y) & [duration, hours]	Dominant underlying rock type	Imagery sources pre- and post-event (resolution)
1	Wairamarama, Waikato	178	Single event (2017)	74 [24] 146 [48]	2 [24] 10–20 [48]	Sandstone, argillite, mudstone	Pre: Pleiades-1A Post: GeoEye-1 (0.5 m)
2	Retaruke, Whanganui	121	Single event (2018)	160 [24] 172 [48]	80 [24] 30 [48]	Mudstone, sandstone, volcanic ash beds	Pre: Pleiades-1A Post: Worldview-2 (0.5 m)
3	Southern Hawke's Bay	175	Single event (2011)	337 [24] 618 [48] 658 [96]	40 [24] >250 [48] 100–250 [96]	Crushed argillite, limestone, mudstone, sandstone	Pre: Orthorectified aerial photography Post: Worldview-2 (0.4 m)
4	Pohangina, Manawatu	30.8	Multi-temporal (1995–2011)	Multi-event 147–195 [24]	50–60 to >250 [24]	Unconsolidated sandstone	Orthorectified aerial photography (0.4–0.75 m)
5	SE Pahiatua, Manawatu	9.8	Multi-temporal (1997–2011)	Multi-event 102–106 [24]	10–20 [24]	Sandstone and mudstone	Orthorectified aerial photography (0.4–0.75 m)
6	Mangamaire, Manawatu	9.4	Multi-temporal (1997–2011)	Multi-event 102–106 [24]	10–20 [24]	Mudstone	Orthorectified aerial photography (0.4–0.75 m)

[a] Average recurrence intervals (ARI) obtained from NIWA's High Intensity Rainfall Design System (HIRDS) v4.

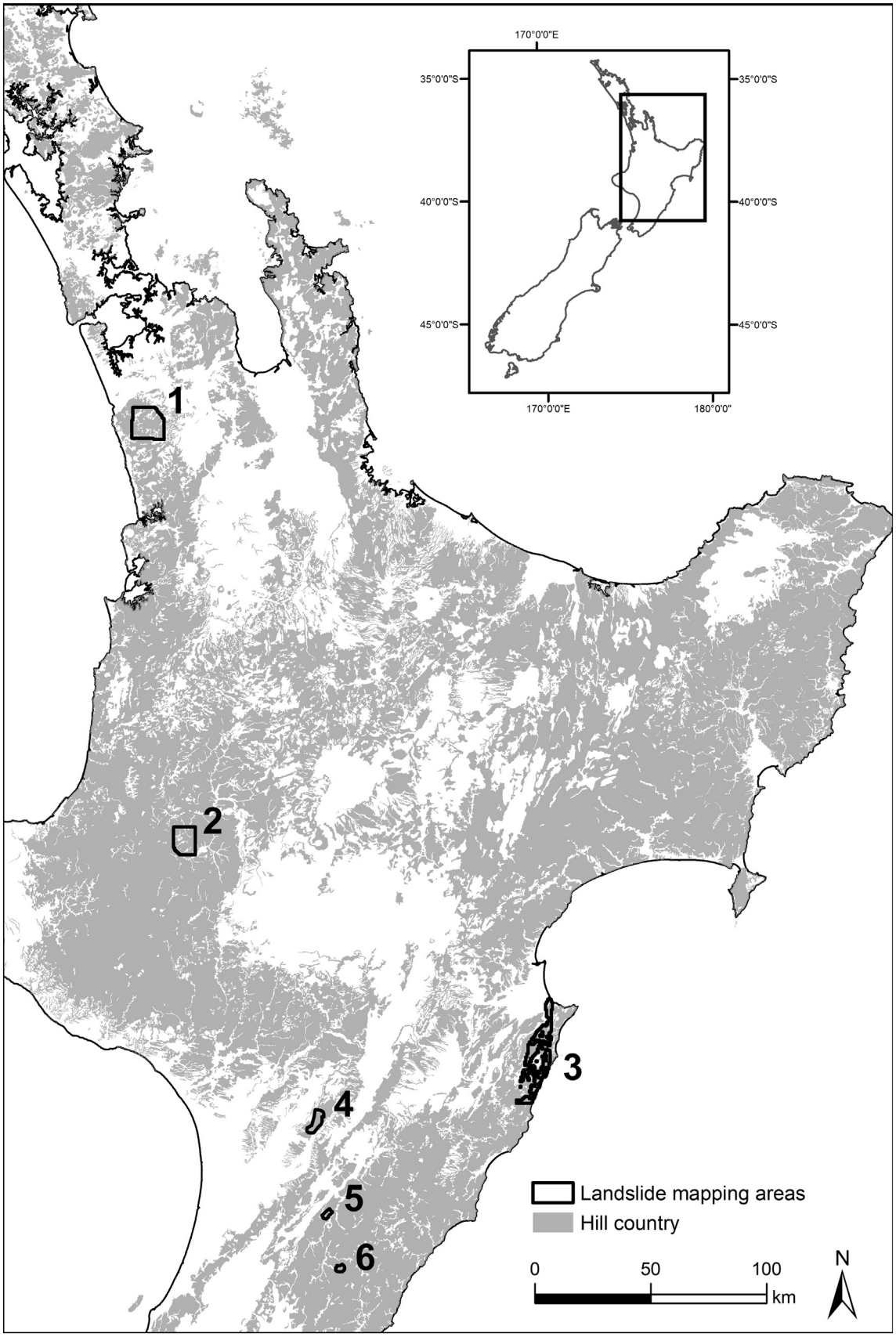

Fig. 1. Location of six study areas for shallow landslide mapping on the North Island of New Zealand, and the extent of hill country (after Dymond et al., 2010).

are underlain by a range of Tertiary and Quaternary sedimentary rocks, which in the Whanganui study area are overlain by volcanic ash predating the Taupo eruption c. 1820 years ago (Table 1). The landslide response is largely attributed to a single storm event for three areas, namely Waikato (2017 storm), Whanganui (2018), and Hawke's Bay (2011), which range in size from 121 to 178 km^2 (Table 1). For the other three areas multi-temporal landslide inventories were prepared.

For the three single events, study area selection was initially guided by the landslide impact (i.e. focusing on the area with the greatest number of landslide scars) based on local reports and lower resolution (5 m) imagery. The delineation of specific areas for mapping was then defined by the availability of mostly cloud-free, intersecting pre- and post-storm VHR (≤0.5 m) satellite or aerial imagery. Due to the relatively infrequent capture of VHR imagery over New Zealand, the choice of suitable intersecting images for mapping was constrained. Hence, the imagery selected (Table 1) represents the best available within the shortest interval before the date of the landslide triggering storm, while post-event imagery was available 6–7 months after triggering events. While this may result in some re-working of landslide scars with subsequent rainfall (see Section 2.2), this post-event interval allowed partial revegetation of debris deposits compared with unvegetated scars that aided separation of landslide scars and debris deposits based on differences in reflectance during manual and semi-automated mapping. For Hawke's Bay, partial cloud cover in the available post-event satellite

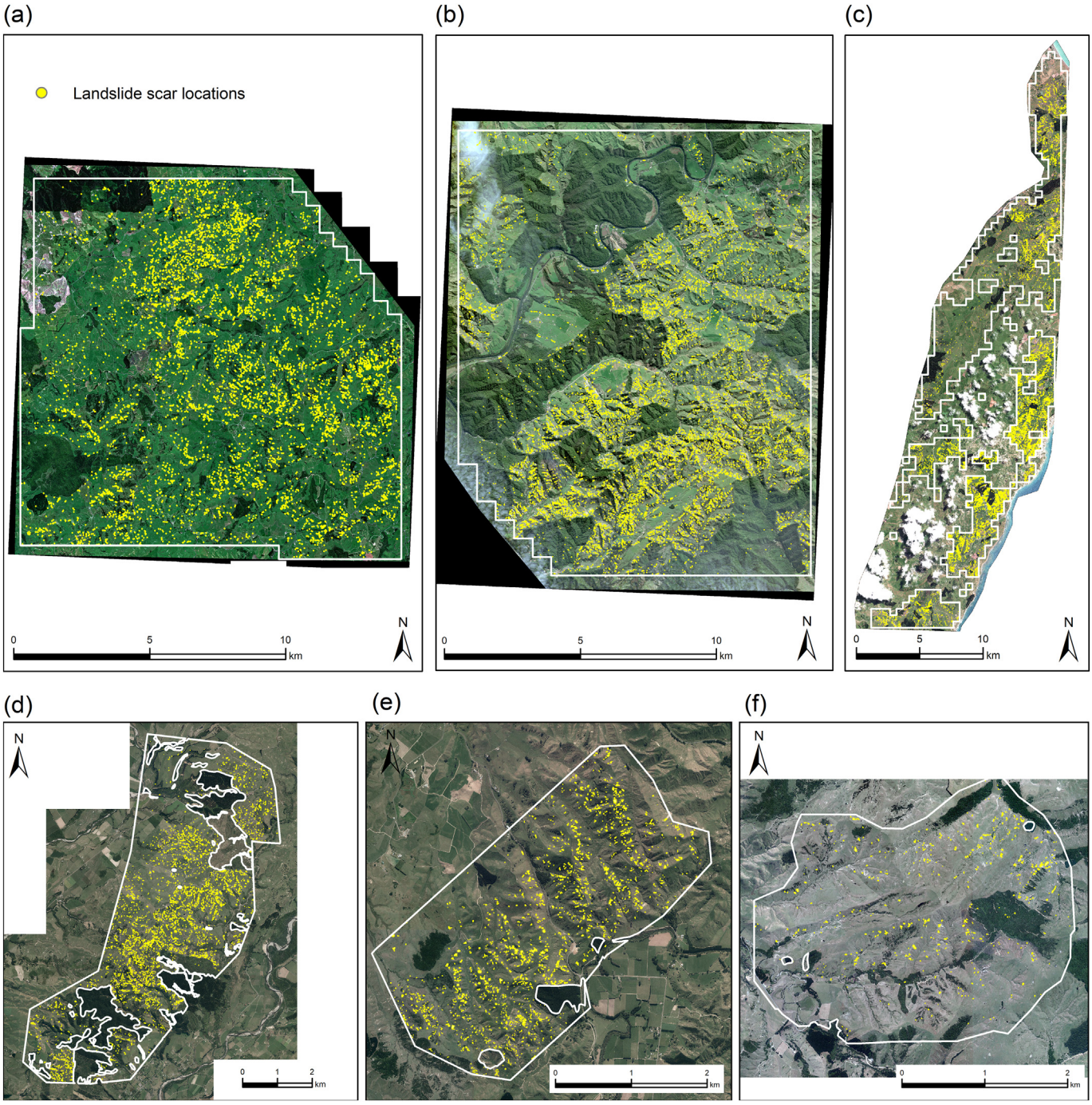

Fig. 2. Mapping extents (boundary line) and shallow landslide scar locations (points or polygons) for a) Waikato, b) Whanganui, and c) Hawke's Bay event-scale study areas and d) Pohangina, e) SE Pahiatua and f) Mangamaire multi-temporal study areas. The Hawke's Bay study area was divided into a grid and those grid cells with partial or complete cloud cover were excluded from mapping.

imagery necessitated exclusion of cloud-affected areas from the wider area subject to manual and OBIA mapping (Fig. 2c).

Study areas 4–6 were originally chosen to represent hill country terrain dominated by different geological parent material and erosion process susceptibility (Dymond et al., 2013). Landslide inventories from these study areas comprise multi-temporal records based on manual mapping from aerial photographs. Data from two of these study areas (Pohangina and SE Pahiatua) were previously reported by Betts et al. (2017) who investigated the effect of differences in underlying geology on the landslide-slope relationship. Subsequently, the size of these study areas was increased with further mapping and a third study area added (Mangamaire) to produce an expanded shallow landslide database. Landslide mapping for the Pohangina, SE Pahiatua, and Mangamaire study areas spans successive aerial photographs captured in the 1990s, 2004, and 2011 (Betts et al., 2017) and provides information on spatial patterns in shallow landslide occurrence for which concurrent land cover data is also available. Landslide data reported by Betts et al. (2017) from earlier aerial photography captured in the 1940s and 1970s was not included due to uncertainty related to the extent of land cover changes prior to commencement of national land cover mapping in 1996. We also exclude areas of exotic forest as these account for land cover change between 1996 and 2012 in the multi-temporal study areas. This ensured that mapped landslides (and non-landslide locations used in modelling) could be related to land cover present over the duration of the multi-temporal study period.

2.2. Storm rainfall and imagery

Storm rainfall data is available from recording sites within or near the three study areas covering event-scale landslide responses (Table 1). While the individual storms that triggered the landslide responses have been confirmed, there is potential for some re-working or the limited formation of new scars with post-event rainfall during the intervals between the storm events and post-event imagery. For comparison, we report rainfall totals for the main landslide-triggering storm event alongside other high magnitude rainfall amounts for each measurement interval.

For the Waikato study area, pre-event Pleiades-1A imagery was available for 4 January 2017. The landslide triggering storm event occurred over 4–5 April 2017 and totalled 146 mm based on Waikato Regional Council's Wairamarama gauge that is located within the study area. The next highest 2-day rainfall total of 99 mm occurred just over one month later and potentially contributed to some re-working of landslide scars and debris deposits. However, this subsequent event falls below the estimated landslide triggering threshold of 125–200 mm in 48 h reported by Basher et al. (2020). Post-event GeoEye-1 imagery was captured on 18 October 2017.

Pre-event Pleiades-1A imagery was available for 4 March 2018 immediately before the storm event that triggered landslides across the Whanganui study area. The total storm rainfall based on the nearest rain gauge located 9 km from the centre of the study area was 172 mm over 7/8 March 2018 (Lower Retaruke CWS, NIWA network no. C95012), of which 160 mm occurred in 1 day. Following the storm, the next highest daily rainfall total of 49 mm occurred 1 month later. Post-event, Worldview-2 imagery was available over the area for 10 September 2018.

The storm event in southern Hawke's Bay had the largest impact on an area near the coast between Cape Kidnappers and Porangahau where rainfall ranged from ~200 to 650 mm across the 175-km^2 study area during 25–28 April 2011 (Jones et al., 2011). The Hawke's Bay Regional Council's Waipoapoa gauge, located near the centre of the study area, recorded total storm rainfall over the 4-day period of 658 mm, with highest 1- and 2-day totals of 337 and 618 mm, respectively. Regional aerial photography from the 2010/11 summer period was available before the event, while post-event Worldview-2 imagery was captured on 1 December 2011. An earthquake (M 4.5 Richter magnitude) occurred during the storm but was below the minimum magnitude typically required to produce a landslide response (Jones et al., 2011). During the interval after the April storm, another event occurred with a 2-day total of 193 mm (22–23 July). This may have contributed some additional landslide scars and to re-working of existing scars.

Storm event rainfall was compiled by Betts et al. (2017) for the Pohangina and SE Pahiatua/Mangamaire study areas. The range in reported maximum daily totals was 147–195 and 102–106 mm for the Pohangina and SE Pahiatua/Mangamaire study areas, respectively, during our multi-temporal study period. The largest number of landslides in the multi-temporal inventory was associated with a single storm event in 2004 (Betts et al., 2017).

2.3. Manual landslide mapping

Manual mapping comprised both point and area-based representation of individual landslide source areas (scars). The size of study areas 1–3 (Table 1) and the large number of scars present meant it was not practical to manually delineate all scar areas. Therefore, we applied point-based manual mapping for these areas (cf. Petschko et al., 2015) that involved placing a single point at the centre of individual landslide source areas. This mapping excluded landslides detected in pre-event imagery. In contrast, manual delineation of all landslide scar areas was previously completed for the smaller multi-temporal study areas 4–6 (Table 1) as reported by Betts et al. (2017).

We selected a subset of landslide scars for manual area mapping in study areas 1–3 using a random grid sampling approach. These data were used to evaluate the classification performance of semi-automated mapping. The procedure involved dividing each study area into a 0.5 × 0.5 km grid, randomly selecting 10% of the grid squares, and mapping all scars present within the selected squares (Fig. 3). All manual landslide delineation was performed by the same person and followed the procedure outlined by Betts et al. (2017) in which landslide margins were delineated using ArcGIS 10.5 at on-screen scales ranging 1:600 to 1:1250. Landslides were divided into scar and debris deposit areas based on differences in reflectance (Betts et al., 2017), but only scar data are used for landside susceptibility modelling.

2.4. OBIA classification

A semi-automated mapping procedure was applied as an alternative approach to identify and classify landslide scar features for study areas 1–3. The primary advantage of this approach over manual delineation of scars is that it enables rapid mapping using a set of defined rules that ensures consistency across complete study areas. The OBIA method combines image processing and GIS functions to delineate and classify homogenous objects (Blaschke, 2010; Blaschke et al., 2014). We used Trimble's eCognition software and employed a knowledge-based ruleset (Hölbling et al., 2016).

Fig. 4 illustrates the OBIA procedure used for detecting and classifying shallow landslides. Three distinct classifications were performed to delineate landslides. First, pre-existing erosion features were removed by mapping bare ground in the pre-event VHR multispectral (MS) imagery, except for the Hawke's Bay study area where available pre-event aerial photography was limited to RGB. Next, farm tracks and roads were classified in the post-event imagery, as these are often associated with areas of bare ground and are a potential source of error. Finally, landslide scars were identified in the post-event imagery, with false positives removed by overlaying pre-existing bare ground, roads, and tracks.

For each of the three rule-based classifications, image objects were created using the multiresolution segmentation algorithm based on spectral information of the imagery. Scale, shape, and compactness parameters were optimised based on the classification objective and image resolution. For example, roads and farm tracks are typically long, linear features in the landscape, so the compactness parameter

Fig. 3. Example of manually delineated shallow landslide scars from the Hawke's Bay study area. Post-storm event landslide scars were identified based on comparison with pre-event imagery.

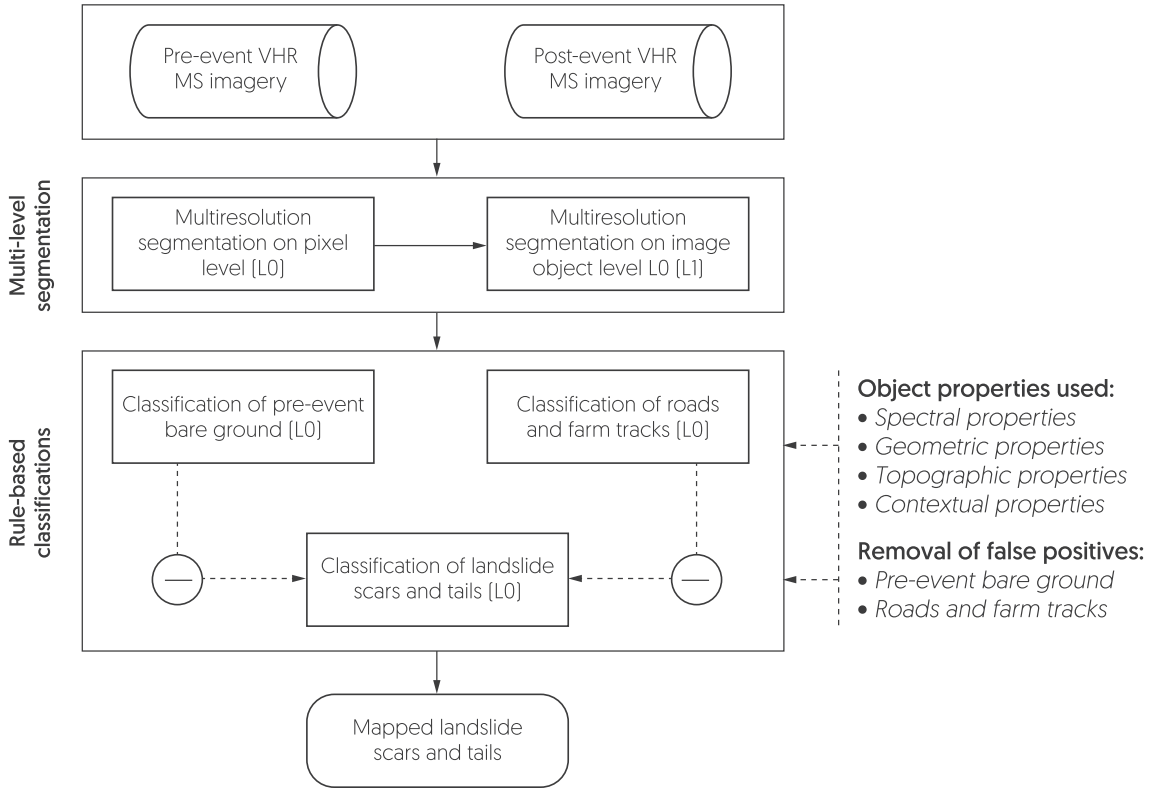

Fig. 4. Workflow for OBIA mapping of shallow landslide scars across study areas 1–3.

was reduced to allow objects to be elongated. Following the creation of image objects, geometric and topographic properties (slope < 20°) were used in conjunction with the spectral properties of objects (high brightness and low normalised difference vegetation index (NDVI) values) to classify roads and tracks. For Hawke's Bay, NDVI could not be calculated for the pre-event imagery, thus only brightness was used.

To classify both pre-existing bare ground and landslides associated with the storm events, two object levels were created using a multi-scaled segmentation approach, which organizes the objects into a hierarchy (Blaschke et al., 2014). The level 0 objects (L0) were created at the pixel level and aim to represent small landslide scars or components of larger landslides. The level 1 objects (L1) were generated through a second multi-resolution segmentation based on the L0 objects. This generally resulted in the delineation of individual paddocks/land parcels, the borders of which were shared with the underlying L0 objects (Blaschke et al., 2014). The advantage of this approach is that the description of low-level objects is enhanced through consideration of their local environmental setting. This was achieved by calculating the mean difference in brightness and NDVI between L0 and L1. This method increases local contrast and allows detection of landslide scars that would otherwise be unfeasible using single thresholds for complete study areas.

The final step sought to differentiate landslide scars from debris deposits. Due to the exposed subsoil or rock surface, scars are generally brighter objects with less variance in the spectral values than debris deposits. The deposits typically have lower brightness values, higher NDVI values, and greater variance that reflects intermixed grass cover and bare soil. Additionally, they are situated below scars. The rule-set thus follows a 2-step approach in which scars are initially classified, followed by debris deposits using the following criteria: a debris deposit must be an object that borders a scar and has 1) lower mean elevation than the adjacent scar object and 2) a low mean and high standard deviation NDVI. This procedure was implemented as a loop to allow deposits to grow down a slope across multiple L0 objects.

The classification performance of OBIA-based mapping was evaluated by comparison with the random grid-based sample of manually mapped landslide scars from study areas 1–3. Accuracy was assessed according to the number of manually mapped landslide scars that intersect with OBIA-mapped scars. Producer's and user's accuracies were computed where the producer's accuracy refers to the number of manual scars intersecting OBIA scars relative to the total number of manually mapped scars. User's accuracy equates to the number of manual scars intersecting OBIA scars relative to the total number of OBIA-mapped scars and provides a measure of reliability. In making this comparison, we also recognise the challenge associated with using reference data from manual mapping that depends on expert judgement (Hölbling et al., 2015).

2.5. Explanatory variables

Explanatory variables for landslide susceptibility modelling were obtained from national datasets for elevation (15-m digital elevation model), rock type mapped at 1:63,360 scale with some areas subsequently remapped at 1:50,000 scale (NZ Land Resources Inventory, Newsome et al., 2008), and land cover from thematic classification of satellite imagery (NZ Land Cover Database, LCDB, from 2012 and 2018) with a minimum mapping unit of 1 ha. Slope angle and flow accumulation (continuous variables), aspect (categorical variable with 8 classes), profile (categorical: convex, concave, planar) and planform curvature (categorical: convergent, divergent, planar) were also derived from the DEM (Table 2). These input variables were selected because a) there is a physical basis for how each variable may influence landslide susceptibility (outlined below), b) model inputs can be derived for all study areas, and c) use of national datasets does not impose a potential spatial constraint on predictions of landslide susceptibility required for land management. Instead, limits on the applicability of susceptibility predictions for different terrain primarily relate to the landslide inventories available for modelling.

Slope is the most widely used explanatory variable in landslide susceptibility modelling (Reichenbach et al., 2018). Slope angle exerts a fundamental physical control through its effect on the shear stress acting on a slope and the resulting shear strength required to maintain stability (Crozier et al., 1980; Crozier, 2010). Several studies in New Zealand have identified variation in slope as the key factor driving spatial patterns in shallow landslide occurrence (Dymond et al., 2006; De Rose, 2013; Betts et al., 2017). Topographic form may influence susceptibility via its effect on soil water movement (Crozier et al., 1980). The inclusion of flow accumulation (upslope contributing area of each pixel) as a model input may reflect patterns in water flow and potential soil saturation (Catani et al., 2013). Calculation of planform curvature allows delineation of convergent zones that concentrate flow and increase the likelihood of soil saturation that could contribute to failure (Catani et al., 2013; Crozier, 2017). Profile curvature represents the rate of change in slope angle in the direction of maximum slope and helps discriminate between erosional versus depositional areas.

Aspect may influence susceptibility via a) differences in solar radiation affecting soil moisture, vegetation growth and soil development (Burnett et al., 2008; van Westen et al., 2008; Inbar et al., 2018), b) bedding angle (Ruff and Czurba, 2008) or c) rainfall direction associated with landslide-triggering storm events (Liu and Shih, 2013). Aspect has been shown to be a factor in the spatial patterns of shallow landslides in New Zealand, where northern aspects were reported as more vulnerable to landslide occurrence (Crozier et al., 1980; Gao and Maro, 2010). This was attributed by Gao and Maro (2010) to greater solar radiation increasing wetting-drying cycles on these slope aspects causing soil cracking that allows more rapid saturation of subsoils during storm events.

Rock type may also influence landslide susceptibility. This occurs indirectly through its effect on soil properties (e.g. texture, permeability) and slope angle rather than directly via bedrock strength given that shallow landslides (typically 0.5–1 m deep) in New Zealand generally occur in regolith (Crozier et al., 1980; Reid and Page, 2002; Betts et al., 2017). Notably, Betts et al. (2017) found landslide density to be independent of rock type based on analysis of historic aerial photographs. However, this analysis was limited to weakly and moderately indurated sandstone and mudstone. Land cover affects susceptibility via tree root reinforcement and alterations to soil moisture resulting from differences in canopy interception, evaporation, transpiration and the formation of preferential flow pathways via root systems (Phillips and Marden, 2005; Ghestem et al., 2011; Stokes et al., 2014). Numerous studies report higher densities of storm-generated landslides for pasture compared to forested terrain in New Zealand (e.g. Reid and Page, 2002; Dymond et al., 2006; Basher, 2013).

Table 2
Topography, land cover (LCDB classes) and rock type (NZ Land Resources Inventory) inputs used in shallow landslide susceptibility modelling across the study areas.

Data	Model inputs
Topography (T)	Slope; Aspect; Flow accumulation; Planform curvature; Profile curvature.
Land cover (LC)	Indigenous Forest; Exotic Forest; High Producing Exotic Grassland; Low Producing Grassland; Deciduous Hardwoods; Mānuka and/or Kānuka; Broadleaf Indigenous Hardwoods.
Rock type (R)	Argillite – crushed; Floodplain alluvium; Argillite; Limestone; Loess; Mudstone or fine siltstone – banded; Mudstone or fine siltstone – jointed; Mudstone or fine siltstone – massive; Ashes older than Taupo pumice; Sandstone or coarse siltstone – massive; Unconsolidated to moderately consolidated clays, silts, sands, tephra & breccias.

We assembled data for the explanatory variables corresponding to landslide presence and absence locations. Spatial data were extracted based on intersection with either a single manually placed point or internal centroid from polygon mapping for each landslide scar. The use of single points to represent scars instead of polygons that may cover multiple raster cells reduces spatial autocorrelation between observations (Petschko et al., 2014; Goetz et al., 2015). Here, a DEM grid cell (225 m^2) would encompass a typical landslide scar. Moreover, the specific choice of point location within a landslide scar polygon was considered by Petschko et al. (2013) and found to have little effect on model predictive performance. Landslide scars mapped with points in study areas 1–3 were buffered to approximate scar area and excluded from selection as non-landslide locations. The buffer distance was set to the radius of a circle defined by the mean landslide scar area from manual mapping for each study area. Non-landslide points were assigned randomly to any location where a landslide was not recorded with a minimum distance between non-landslide points of two times the scar radius.

The random placement of non-landslide points was repeated five times for each study area to provide a measure of variability associated with the selection of point locations. The number of non-landslide points was set to equal the number of mapped landslide scars for each study area (Goetz et al., 2015; Chen et al., 2017; Lombardo and Mai, 2018). The sensitivity to presence/absence ratio was examined by Heckmann et al. (2014) who found that the choice of ratio did not significantly affect predictive performance using logistic regression.

2.6. Statistical modelling

2.6.1. Logistic regression

Logistic regression is a type of generalised linear model (GLM) that uses a logistic function with a binary dependent variable (landslide presence/absence). It relates the probability (P) of landslide scar presence ($Y = 1$) to the spatial explanatory variables ($x_1, x_2, \ldots x_n$) where $\beta_0, \beta_1, \beta_2 \ldots \beta_n$ are fitted constants (Eq. (1)):

$$P(Y=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots \beta_n x_n)}} \quad (1)$$

Logistic regression is sensitive to collinearity between input variables. We assessed potential collinearity using the variance inflation factor (VIF) with the 'usdm' R package (Naimi, 2015) and a VIF value >10 as a threshold for variable exclusion (Conoscenti et al., 2016). The resulting ranges in VIF values for the study areas were 1.01–1.06 and 1.12–1.49 for flow accumulation and slope, respectively.

We select variables for inclusion in the logistic regression analysis based on prior information on how each variable may influence landslide susceptibility and data availability. We do not use automated stepwise variable selection given potential biases in parameter estimates and other issues that may result from using this procedure (Harrell, 2001; Lombardo and Mai, 2018). To test the effect of variable inclusion on predictive performance, we compare variable combinations comprising 1) topography-based inputs only (i.e. slope, aspect, flow accumulation, curvature), 2) topography and land cover, and 3) topography, land cover and rock type using the logistic regression and random forest models.

2.6.2. Random forest

Random forest models (Breiman, 2001) comprise multiple individual classification trees (to create the 'forest') that are combined to make a prediction. Each tree is formed by taking a random sample with replacement for training with the remaining 'Out-of-bag' (OOB) data used to evaluate the classification error as trees are added. The algorithm also draws a random subset of predictor variables for each node in a tree where the variable achieving the best data split is selected for that node. The final prediction of landslide presence/absence is determined by the majority classification result achieved across all trees in the forest. Several studies have applied random forest models for landslide susceptibility analysis and found they performed well when compared with other models (Goetz et al., 2015; Youssef et al., 2016; Pourghasemi and Rahmati, 2018).

We use R package 'randomForest' (Liaw and Wiener, 2002) to implement Breiman and Cutler's random forest algorithm based on Breiman (2001). The number of trees (ntree) and the number of variables randomly sampled for each data-splitting node (mtry) are selected by the user. Catani et al. (2013) tested how increasing tree number affected prediction accuracy based on the OOB classification error. These authors found that the OOB error began to stabilise from 100 trees and selected 200 trees for modelling. Here, we conservatively set the number of trees to 300 to balance run time with accuracy. For mtry, we use the default setting of the square root of n variables (Liaw and Wiener, 2002).

2.6.3. Model evaluation

We compare predictive performance of the logistic regression and random forest models based on landslide inventories derived from a) manual versus semi-automated mapping and b) event versus multi-temporal datasets. Landslide susceptibility models are usually trained and tested on separate datasets. This typically involves splitting data from a given study area, whether by temporal partitioning or via geographic sampling, either randomly or using spatially selected subsamples (Reichenbach et al., 2018). However, the purpose of modelling is the production of landslide susceptibility maps to inform land and hazard management typically over wider areas or regions. Thus, it is also important to understand how models perform when applied to different study areas within an equivalent spatial domain, which, in the present study, relates to New Zealand's hill country terrain. Therefore, in addition to evaluating models within each study area, we test performance of models fitted using multi-temporal data from study areas 4–6 with event data (study areas 1–3), and vice versa where models fitted with event data are tested with the multi-temporal data.

Model predictive performance was assessed using k-fold cross-validation. Landslide presence and absence datasets were randomly shuffled and then split into $k = 5$ folds, where $k - 1$ folds were used for model fitting and each remaining fold was used once for testing. The same k-fold splits were applied to both statistical models to ensure consistency when comparing predictive performance using the same data within each fold. This procedure was repeated for the five sets of randomised non-landslide locations to produce a combined total of 25 data splits for each study area and combination of input variables. When making statistical comparisons between model results, we applied the Bonferroni correction to adjust for multiple testing. This divides the significance level ($\alpha = 0.05$) by the number of tests performed.

We test model sensitivity to sample size by randomly sampling our presence/absence data in equal proportions (cf. Petschko et al., 2014) using sample sizes of $n = 400, 800, 1600, 3200, 6400, 12{,}800$ as well as all data. This analysis focuses on comparing the two statistical models with the combined event (study areas 1–3) versus multi-temporal (study areas 4–6) inventories. While smaller samples may produce more varied and less accurate predictions, they are also less likely to experience issues with either spatial autocorrelation that violates the logistic regression assumption of independent observations or model overfitting (Heckmann et al., 2014). However, the issue of spatial autocorrelation is less problematic when the focus is on predictive performance rather than statistical inference using model coefficients (Brenning, 2005).

Model classification performance (i.e. landslide versus non-landslide) was evaluated using receiver operating characteristic (ROC) curves and calculation of the area under curve (AUC) based on the 25 iterations using the 'ROCR' package (Sing et al., 2005). ROC curves and AUC values are widely used in the landside susceptibility literature to assess model performance (Reichenbach et al., 2018). ROC curves plot

the true positive rate ($TPR = \frac{TP}{TP+FN}$) versus the false positive rate ($FPR = \frac{FP}{TN+FP}$) to give an indication of sensitivity (probability of detection) versus 1 – specificity (specificity refers to the true negative rate) (Conoscenti et al., 2016). An AUC threshold of 0.7 was considered acceptable following Hosmer et al. (2013). An AUC value of 0.5 corresponds to performance no better than a random guess, while an AUC of 1 would indicate perfect classification. Spatial prediction of landslide probability values employed the 'raster' package (Hijmans, 2019). All statistical analyses were completed using R v3.5.1 (R Core Team, 2018).

3. Results

3.1. Shallow landslide characteristics

Shallow landslide data from manual mapping across the six study areas are summarised in Table 3. Landslide scar densities range 43–155 scars km^{-2} for individual storm events compared with 58–214 scars km^{-2} based on the multi-temporal records (Table 3). The landslide scar mapping shows clear spatial patterns in scar locations (Fig. 2). Over short distances (<1 km) there are distinct changes in the spatial density of landslide scars that correspond to land cover, such as pasture versus indigenous forest cover (e.g. Fig. 2b).

The observed event-scale scar densities increase with the maximum estimated average recurrence interval of each storm event. Maximum ARIs of 10–20 (Waikato), 80 (Whanganui) and >250 (Hawke's Bay) years correspond to increasing scar densities of 43, 105 and 155 scars km^{-2}, respectively. Similarly, a large range in scar densities was observed for study areas 4–6 based on multi-temporal data.

Mean and median landslide scar areas range from 95 to 181 m^2 and 50–106 m^2, respectively, across the six study areas. This is based on a random grid-based sub-sample (10% by area) for study areas 1–3 and complete coverage for the smaller study areas 4–6. The probability density for event versus multi-temporal scar areas is plotted in Fig. 5 and shows similar distributions.

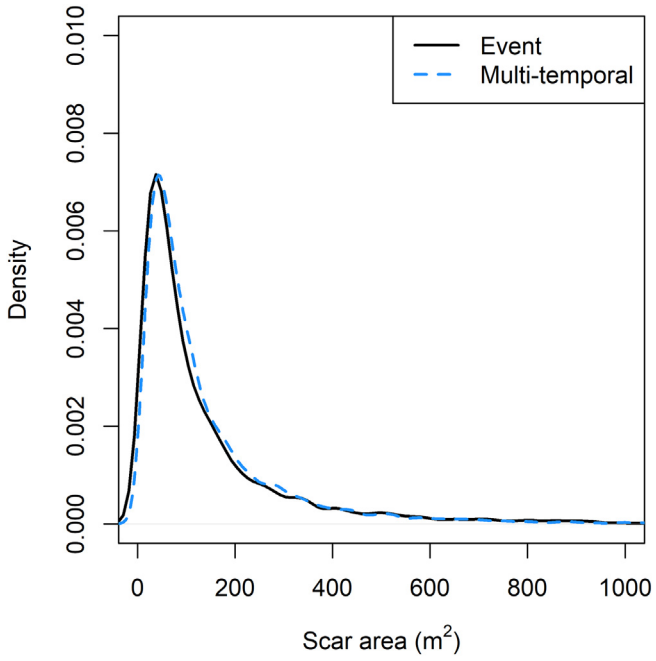

Fig. 5. Probability density plots of shallow landslide scar areas for event (study areas 1–3) versus multi-temporal (study areas 4–6) inventories based on manual delineation of landslide scar areas. For display, the range in scar area on the x axis is limited to 1000 m^2, which excludes 1.2 and 0.5% of the event and multi-temporal datasets, respectively.

3.2. OBIA landslide classification performance

OBIA mapping produced mixed results when compared with manual delineation of landslide scars across 10% of study areas 1–3. OBIA results consistently exceeded the number of landslide scars mapped manually. OBIA performed better for the Whanganui study area (Table 4) where Producer's and User's accuracy were 64 and 55%, respectively. Performance was consistently poorest for Waikato, while there was a large range between the producer's (81%) and user's (46%) accuracies for Hawke's Bay (Table 4). To avoid ambiguity in the comparison of OBIA versus manual mapping results, we did not apply any manual refinement to OBIA outputs. Thus, accuracies demonstrate a base level of performance related to the semi-automated classification procedure only. Further improvements in accuracy would depend on the extent of manual editing applied.

The number of OBIA mapped scars in Hawke's Bay, Waikato and Whanganui exceed manually mapped scars by 77, 38 and 17%, respectively. The higher rate of false positives in Hawke's Bay and Waikato reflects more extensive areas of non-landslide bare ground and rock outcrops that were challenging to discriminate from landslide scars using OBIA. While moderate producer's accuracies were obtained (62–81%), the lower user's accuracies reflect issues in the reliability of OBIA mapping results due to the number of false positives. For Hawke's Bay, the higher false positives may also reflect underestimation of existing bare ground due to the absence of a near-infrared band in pre-event imagery to compute NDVI as part of the classification procedure. Comparison of pre- and post-event imagery enabled the identification and removal of many pre-existing non-landslide features, but minor differences in georectification between images was a partial impediment to this process. Furthermore, the image pairs were spaced several months apart, which introduces challenges resulting from seasonal (e.g. deciduous trees) or land cover changes (e.g. forest harvesting). Temporal proximity of pre- and post-images increases the likelihood that observed spectral changes are associated with the storm rather than other environmental changes but reduces the ability to discriminate scar and debris deposit features.

3.3. Landslide susceptibility

3.3.1. Prediction within study areas

Mean AUC values from cross-validation within study areas range 0.67–0.80 and 0.67–0.81 for logistic regression and random forest

Table 3
Summary of shallow landslide data from manual mapping of the six study areas. The mapped number of landslides for study areas 1–3 excludes scars present in pre-event imagery. The number of mapped scar polygons is based on a random grid-based sub-sample (10% by area) for study areas 1–3 versus complete coverage for the smaller study areas 4–6.

Study area no.	Location	Measurement period	Total no. of scars [point or polygon]	Scar density (scars km^{-2})	No. mapped scar polygons	Mean scar area (m^2)	Median scar area (m^2)
1	Waikato	Single event (2017)	7704	43	1021	95	50
2	Whanganui	Single event (2018)	11,585	105	920	180	106
3	Hawke's Bay	Single event (2011)	27,170	155	1478	181	84
4	Pohangina	Multi-temporal (1995–2011)	4782	155	4782	154	98
5	SE Pahiatua	Multi-temporal (1997–2011)	2096	214	2096	118	62
6	Mangamaire	Multi-temporal (1997–2011)	548	58	548	120	71

Table 4
Producer's and user's accuracy based on comparison of the number of intersecting landslide scars between manual and OBIA mapping for the three event-scale study areas.

Study area	Count manual	Count OBIA	Count manual intersecting OBIA	Producer's accuracy (%)	User's accuracy (%)
1. Waikato	1021	1404	629	62	45
2. Whanganui	920	1074	593	64	55
3. Hawke's Bay	1478	2623	1194	81	46

models, respectively, using manual inventories with different variable inputs (Table 5). The addition of explanatory variables improved model predictive performance based on cross-validation results. The largest increase in AUC occurred with the addition of land cover in combination with topographic variables (i.e. slope, aspect, flow accumulation, curvature). This difference in AUC was significant for both models (Wilcoxon signed-rank test for dependent samples, $p < 0.005$, with the Bonferroni correction), while the further addition of rock type produced a small but significant increase ($p < 0.005$) in all but one case (Whanganui study area). Variation in AUC values related to repeated random spatial selection of non-landslide locations was minor and equated to a maximum range of 0.01 across all variable combinations and inventories.

The random forest model produced slightly better prediction performance compared with logistic regression when tested within the same study area. This small difference between models was significant for all study area inventories except Waikato (Wilcoxon signed-rank test, $p < 0.01$). While differences were minor, based on cross-validation results, the landslide susceptibility maps produced by each model differ more markedly in terms of the probability values and in the extent of higher susceptibility areas (Fig. 6). This difference is evident in histograms comparing pixel values from logistic regression versus random forest predictions (Fig. 7). Random forest produces greater discrimination between stable versus unstable land, whereas logistic regression probability values are closer to 0.5, which has been shown to correspond to higher uncertainties compared with values nearer 0 or 1 (Guzzetti et al., 2006; Rossi et al., 2010).

Model predictive performance was similar when comparing individual event, combined event (Fig. 8a), and multi-temporal (Fig. 8b) datasets using cross-validation with manual data for each study area (Table 5). The range in mean AUC values (0.71–0.77) for the multi-temporal landslide inventory (combined data from study areas 4–6) falls within the range of AUCs obtained from the individual and combined storm event inventories (0.67–0.81). The lack of clear improvement in predictive performance for the multi-temporal compared to event-based inventories may reflect the density of shallow landslides within the event study areas. Increasing event landslide density (range 43–155 scars km^{-2}) generally corresponds with increasing mean event AUC values in the sequence of Waikato, Whanganui and Hawke's Bay study areas.

Comparison of model performance using all input variables with increasing sample size shows an increase in AUC and decrease in interquartile range (IQR) (Fig. 9), as would be expected. The increase in mean (and median) AUC with sample size was similar between logistic regression and random forest models, which ranged between 0.72 and 0.77 (0.72–0.77) and 0.70–0.79 (0.71–0.79), respectively. The decrease in IQR spanned 0.09–0.003 and 0.12–0.01 across both models for the event versus multi-temporal inventories, respectively. Above a sample size of $n = 6400$, predictive performance was consistent with mean and median AUC > 0.76 and IQR < 0.02 for both models and inventories.

Model predictions based on the combined event versus multi-temporal inventories were generally consistent irrespective of sample size (Fig. 9). This comparison based on equal sample sizes indicates that differences in performance when using all data (Table 5) are not simply a reflection of a difference in the amount of data available for model fitting. Notably, the largest sample of $n = 12{,}800$ corresponds to 14 and 87% of the combined event and multi-temporal inventories, respectively. This shows that despite having access to a larger proportion of the available inventory for fitting, the multi-temporal-based models did not outperform the combined event models in cross-validation.

3.3.2. Manual vs. OBIA landslide inventories

Mean AUC values associated with susceptibility models based on manual point mapping consistently exceed OBIA-based results (Table 5). The differences in AUC between manual versus OBIA averaged 0.06, 0.05, and 0.08 for Waikato, Whanganui, and Hawke's Bay, respectively. However, comparison of susceptibility maps based on manual versus OBIA landslide inventories show that patterns in susceptibility are generally similar. As an example, Fig. 10 contains susceptibility maps based on logistic regression with topography, land cover and rock type inputs. Areas exhibiting higher probability values are reasonably consistent between the manual and OBIA based maps.

The largest difference in susceptibility occurs in the Waikato study area where a high number of false positives in the north-west corner resulted in higher predicted susceptibility based on the OBIA compared with the manual inventory (Fig. 10). Histograms of modelled pixel probabilities for each study area show susceptibility results based on OBIA compared to manual inventories have fewer values close to 0 or 1 and more values close to 0.5 (Fig. 11). This mostly resulted in underprediction of stable areas compared to manual inventories, particularly for Waikato and Hawke's Bay, where there were a higher number of OBIA false positives.

3.3.3. Prediction between study areas

The range in mean AUC values for models fitted with event data and tested with the multi-temporal data was 0.66–0.73 across both model types (Table 6). Conversely, the range in AUCs for models fitted with the multi-temporal data and tested with event data was 0.62–0.79. When individual event inventories are combined and used to fit the susceptibility models, predictive performance based on testing with the multi-temporal data (Fig. 8c) produced mean AUC values of 0.68–0.74.

Table 5
Comparison of mean AUC values from cross-validation within study areas for logistic regression (LR) and random forest (RF) models with different input variables using manual versus OBIA inventories for individual events as well as manual data for the combined events (study areas 1–3) and multi-temporal (study areas 4–6) landslide inventories.

Model	Input variables	Waikato		Whanganui		Hawke's bay		Combined event	Multi-temporal
		Manual	OBIA	Manual	OBIA	Manual	OBIA	Manual	Manual
LR	T	0.67	0.63	0.68	0.63	0.74	0.66	0.69	0.71
	T + LC	0.72	0.65	0.79	0.74	0.77	0.69	0.76	0.75
	T + LC + R	0.73	0.66	0.80	0.75	0.79	0.71	0.77	0.77
RF	T	0.67	0.63	0.69	0.63	0.75	0.67	0.70	0.72
	T + LC	0.72	0.66	0.79	0.74	0.77	0.70	0.76	0.75
	T + LC + R	0.73	0.67	0.80	0.75	0.81	0.73	0.78	0.77

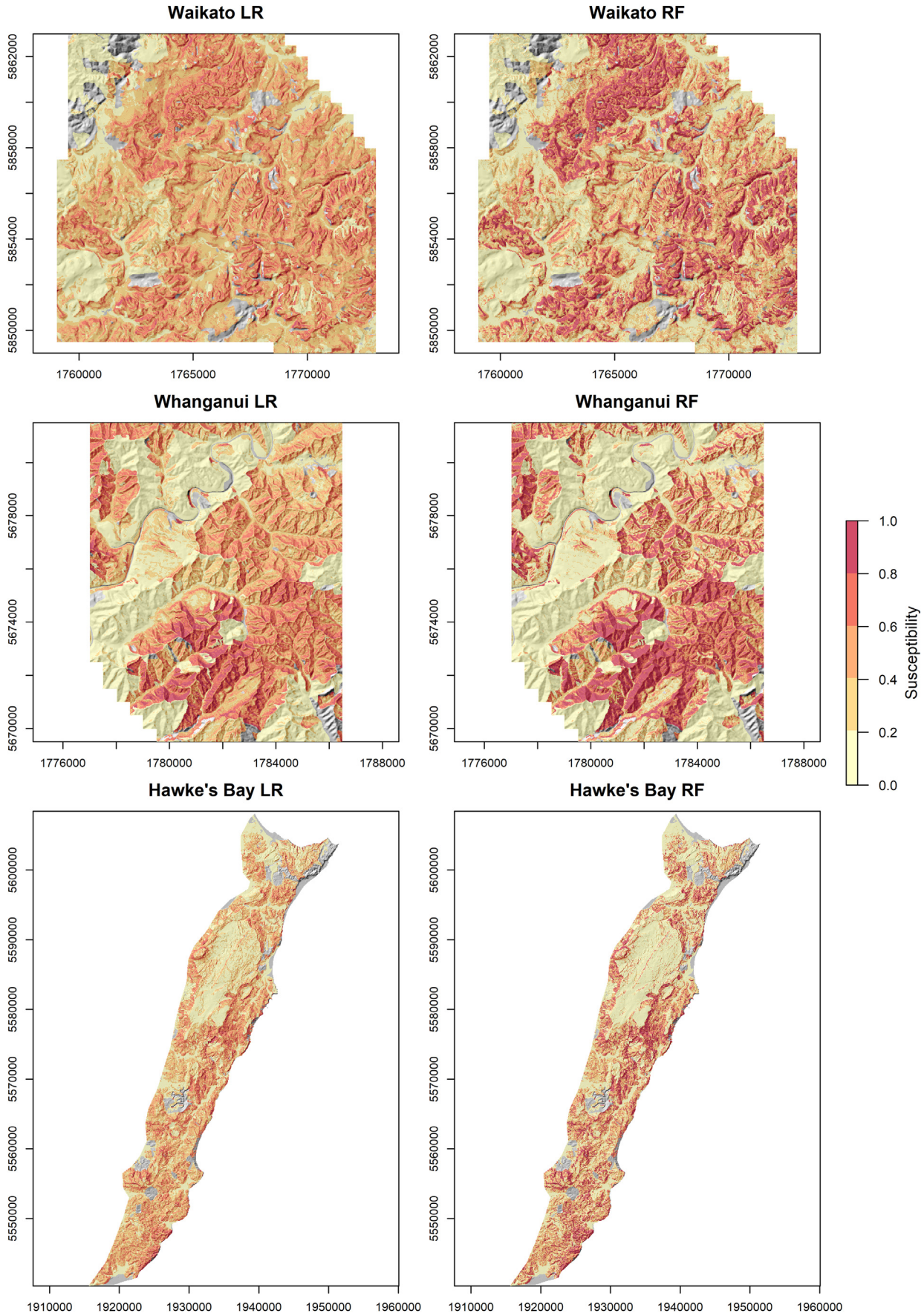

Fig. 6. Comparison of landslide susceptibility based on the means of logistic regression (LR) versus random forest (RF) model predictions ($n = 25$ iterations each) with topographic, land cover and rock type inputs and manual inventory data for the Waikato, Whanganui, and Hawke's Bay event study areas.

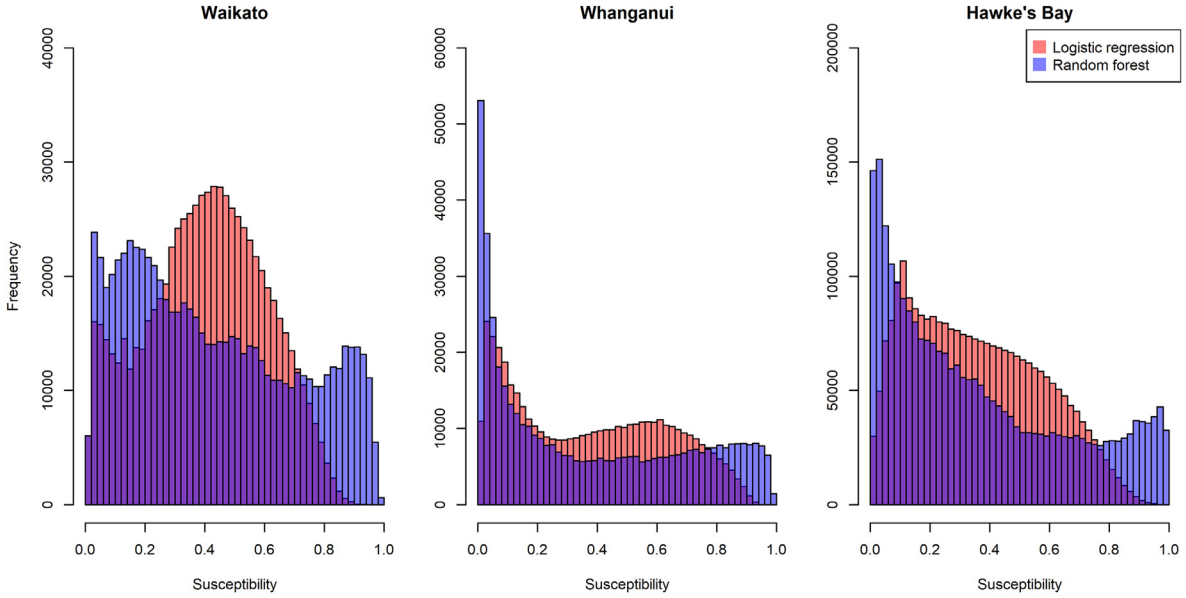

Fig. 7. Histograms of mean modelled pixel susceptibility values for the Waikato, Whanganui and Hawke's Bay study areas comparing logistic regression versus random forest model predictions shown in Fig. 6.

This compares with 0.65–0.74 from fitting with multi-temporal data and testing with combined event data (Fig. 8d). The logistic regression model outperformed random forest when fitting and testing with data from different study areas in 22 out of 24 comparisons (Table 6). This contrasts with the minor outperformance by random forest over logistic regression using cross-validation within a single study area (Table 5).

Increasing the number of input variables did not necessarily improve performance when testing with data from different study areas. The combination of land cover with topographic variables improved model performance, but this decreased with the addition of rock type (Table 6; Fig. 8), in contrast to cross-validation results from within a single study area. Model fitting with multi-temporal data tended to

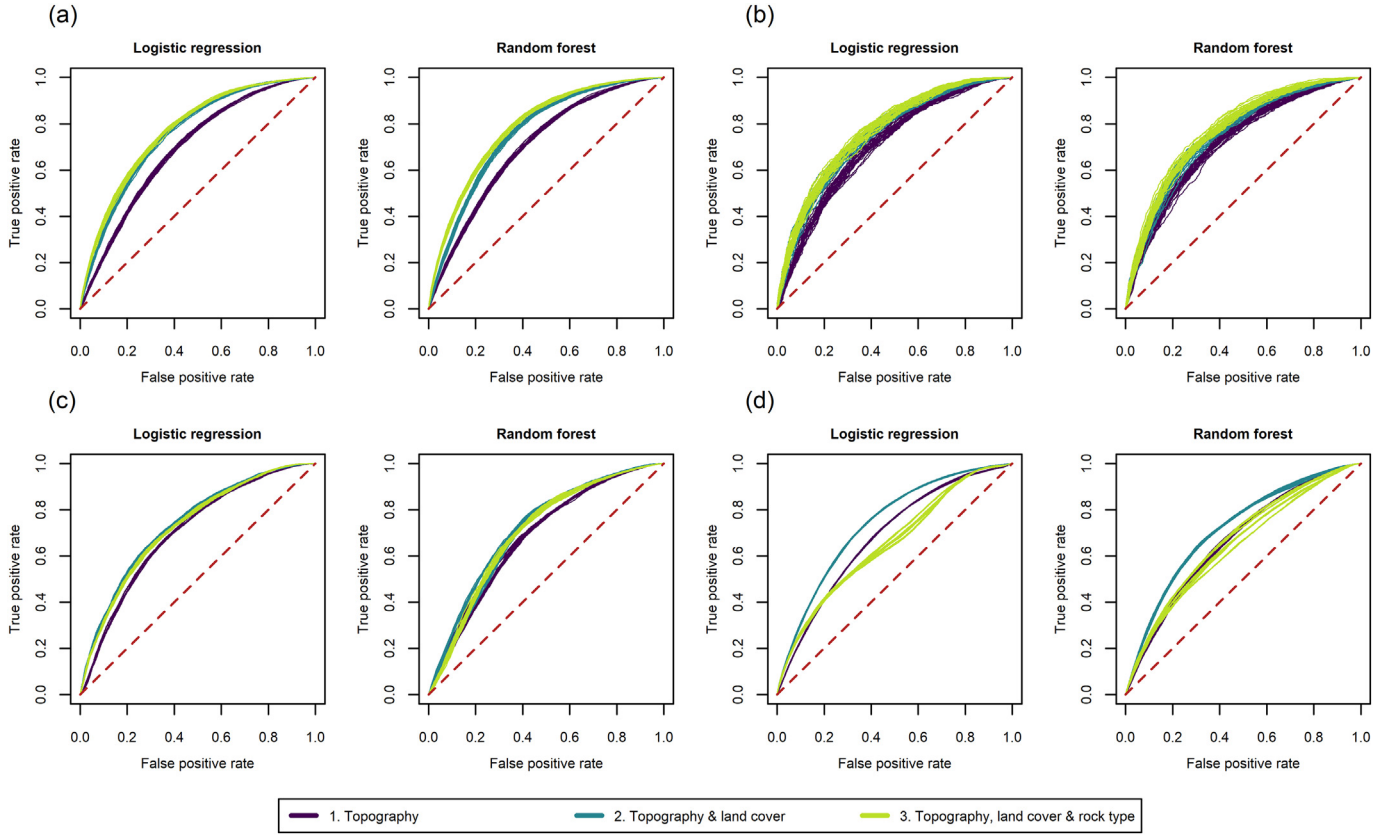

Fig. 8. ROC curves for logistic regression and random forest models. Top row: models fitted and tested using cross-validation with data from within the same study area comprising a) combined event data from study areas 1–3 versus b) multi-temporal data from study areas 4–6. Bottom row: models fitted and tested in different study area comprising c) fitted with combined event data and tested with multi-temporal data versus d) fitted with multi-temporal data and tested with combined event data.

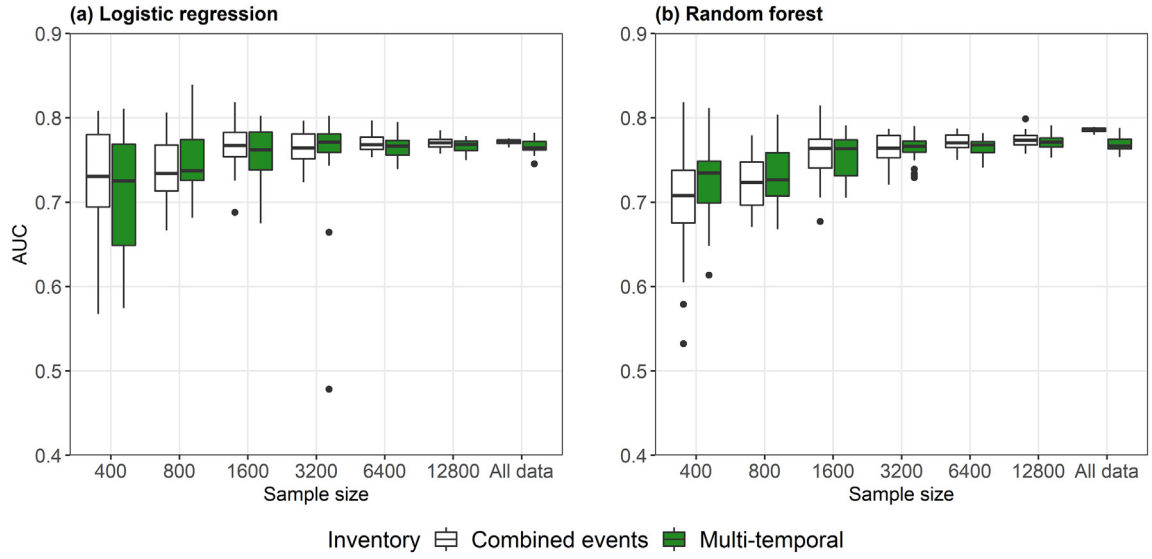

Fig. 9. Comparison of a) logistic regression and b) random forest model predictive performance based on the combined events (study areas 1–3) versus multi-temporal (study areas 4–6) landslide inventories using different sample sizes as well as all data with topographic, land cover, and rock type inputs.

outperform fitting with event data using the combination of topographic and land cover input variables that produced the highest AUCs, but this was reversed with inclusion of rock type (Table 6). The proportion of presence/absence data available for testing (ranging 20–97%) was reduced with the addition of rock type due to the absence of equivalent rock types in both fitting and testing datasets. Combining event inventories increased the range of rock types available for testing with the multi-temporal data to 97%, whereas the multi-temporal inventory is less representative of the range of rock types and this limited the data available for testing to 20–70% of the event presence/absence data. By comparison, the inclusion of land cover produced only a minor reduction in the corresponding data available for testing (range 86–100%).

4. Discussion

The shallow landslide inventories assembled here are consistent with data from other landslide-triggering rainfall events in New Zealand. For instance, Crozier (2005) reported a median landslide density of 30 landslides km^{-2} and a range of 6–478 landslides km^{-2} based on a collation of landslide-triggering rainfall events. These data compare with densities of 43–155 scars km^{-2} for storm events recorded in the present study. De Rose (2013) found that the mean and median scar areas ranged between 46–129 and 32–98 m^2, respectively, based on a historic reconstruction of landslide events in Wairarapa hill country on the North Island. These scar areas are comparable to values obtained for the six study areas presented here (Table 3), where the ranges in mean and median scar areas were 95–181 m^2 and 50–106 m^2, respectively. The small size of scars formed by these shallow landslides underscores the need for high-resolution imagery to ensure detection of scar features. Imagery with a pixel resolution of 10 m or greater risks significantly under-estimating the spatial occurrence of shallow landslide scars.

The OBIA method for mapping landslide scars was previously applied to the SE Pahiatua study area and compared with manual mapping (Hölbling et al., 2016). For natural-colour images, these authors reported landslide detection rates using OBIA of 83 and 93% based on the intersection of landslide scars. This compares with 62–81% detection rates for study areas 1–3 in the present study. Our application of OBIA for landslide classification represents a significant increase in the size of mapping area from 10.1 km^2 in Hölbling et al. (2016) to 121–178 km^2. The increase in study area size is associated with increased diversity in land cover, variations in soil colour within landslide scars/deposits, and the presence of rocks that complicated spectral-based landslide classification, which may in part account for the lower detection rates and higher number of false positives. In addition, no manual refinement was applied in the present study, whereas this formed part of the procedure applied by Hölbling et al. (2016). Semi-automated mapping was also previously performed for the wider Hawke's Bay region affected by the April 2011 storm event using lower resolution (5 m) RapidEye imagery (Jones et al., 2011). However, the resolution was insufficient to allow separation of scars and debris deposits and resulted in a minimum detection size of 100 m^2, which exceeds the median scar size of 84 m^2 for this study area based on manual mapping from VHR imagery.

Rainfall triggering thresholds were evidently exceeded across most of the event study areas given the distribution and number of landslides observed (Fig. 2). Using historic records of landslide events and rainfall data in New Zealand, Glade (1998) estimated that events with >120 mm in 24 h were very likely to trigger landslides, while 125–200 mm over 48 h has been identified as a landslide response threshold (Reid and Page, 2002; Basher et al., 2020). Both the Whanganui and Hawke's Bay events exceed the 24 h threshold, whereas Waikato exceeded the minimum 48-h threshold value. Event landslide scar density (43–155 scars km^{-2}) increases with the maximum ARI (10–20 to >250 years) and generally corresponds with improving model predictive performance (Table 5). This is consistent with an increasing proportion of slopes failing as storm magnitude increases, which enhances the model's ability to discriminate stable and unstable land. It also fits with the conceptual zonation of landslide response according to storm rainfall, as per Crozier (2017)'s cell model for regional landslide triggering events. However, in the present study, these zones correspond to areas mapped within discrete storm events rather than a gradient within a single event.

Model performance in the present study was improved compared to the few previous studies of landslide susceptibility in New Zealand. For example, the heuristic shallow landslide susceptibility model developed by Dymond et al. (2006) following a large storm in February 2004 over the Manawatu region on the North Island was evaluated using SPOT5 imagery with 10-m resolution by applying unsupervised classification. This approach could not separate scars from debris deposits and instead relied on an algorithm to assign pixels to scars. The authors reported

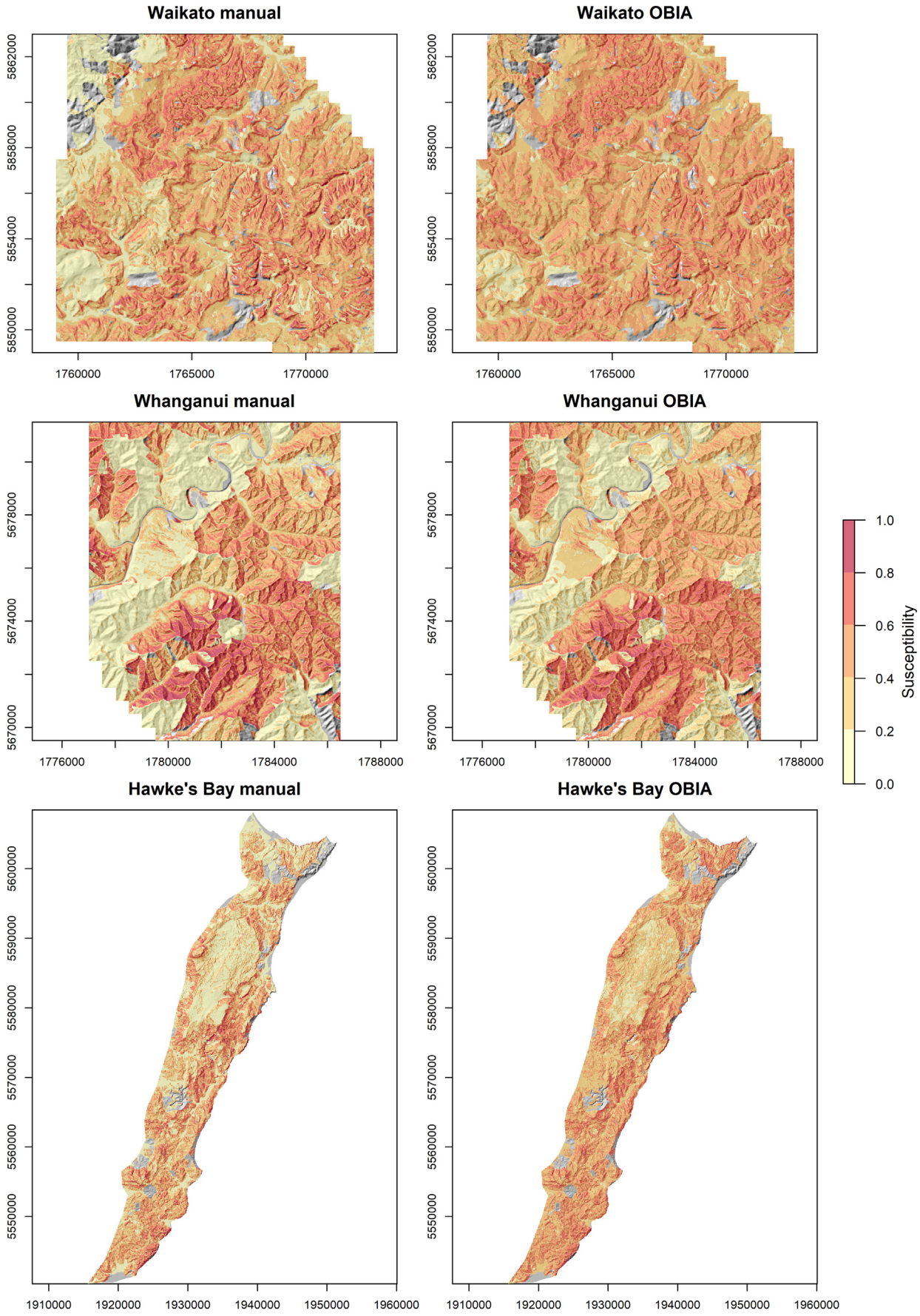

Fig. 10. Comparison of shallow landslide susceptibility maps based on the mean of logistic regression model predictions ($n = 25$ iterations) with topographic, land cover, and rock type inputs for manual versus OBIA derived landslide inventories.

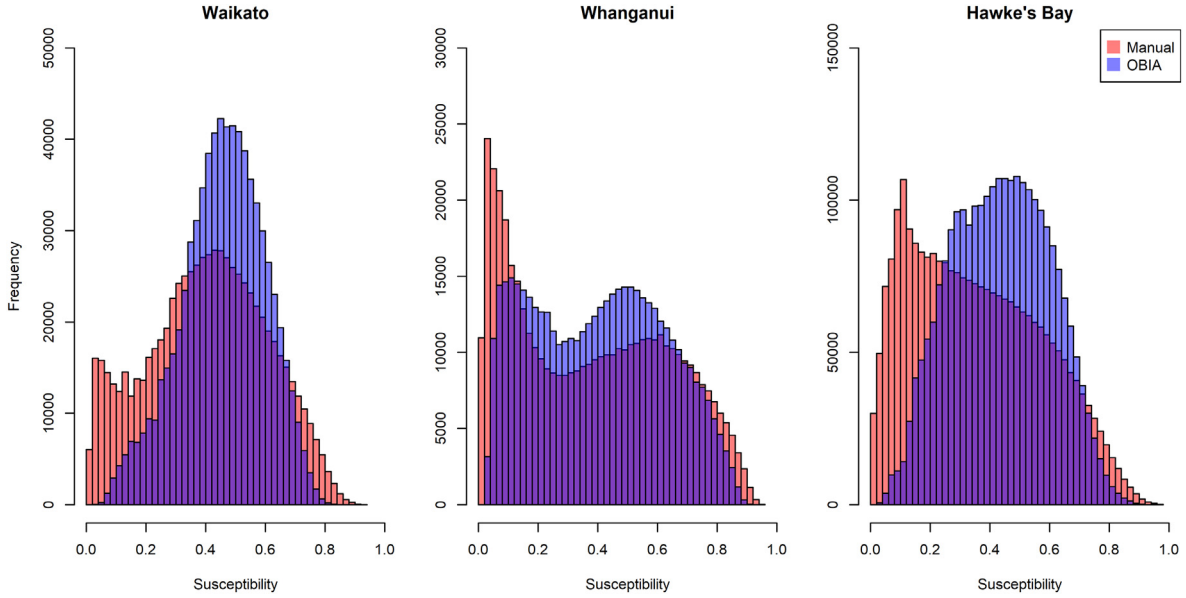

Fig. 11. Histograms of mean modelled pixel susceptibility values for the Waikato, Whanganui and Hawke's Bay study areas comparing logistic regression model predictions based on manual versus OBIA landslide inventories shown in Fig. 10.

only 26% of pixels classified as landslides occurred on land classed as susceptible, while an alternative accuracy assessment based on hillsides (assumed to be susceptible if >25% by area was susceptible) increased accuracy to 58%. This compares with cross-validation accuracies in the present study that range 67–73 and 67–74% from the logistic regression and random forest models, respectively, using all explanatory variables for different inventories derived from manual mapping using higher resolution imagery (0.4–0.75 m). Dymond et al. (2006) noted that use of a slope threshold (range 24–28°) was a poor approximation for susceptible terrain. They also considered all land with woody vegetation as non-susceptible. In contrast, the statistical modelling approach adopted here does not assume that there is no susceptibility below a fixed slope threshold or for slopes under woody vegetation. The latter can experience landslides but at much lower spatial densities than grassland areas (Glade, 1998; Dymond et al., 2006).

Other quantitative studies of landslide susceptibility in New Zealand do not explicitly focus on shallow landslides in hill country terrain. Schicker and Moon (2012) employed logistic regression and weights of evidence methods with an inventory of predominantly large landslides and reported AUC values of 0.71–0.75. Kritikos and Davies (2015) examined rain-triggered shallow landslides in the mountainous western Southern Alps of New Zealand, which contrasts with hill country that forms the focus of the present study. These authors reported AUC values ranging 0.71–0.73. By comparison, mean AUC values from cross-validation using all explanatory variables in the present study showed modest improvement based on combined event and multi-temporal manual inventories (range 0.77–0.78, Table 5), with a range in AUC values for individual event study areas of 0.73–0.81. The future availability of regional LiDAR-derived DEMs in New Zealand should enable further improvement in susceptibility modelling through better representation of topographic inputs.

Model comparison showed only minor improvement in predictive performance by random forest over logistic regression, whereas there were notable differences in the landslide susceptibility maps (Fig. 6) and histograms of pixel probability values (Fig. 7). Susceptibility maps based on logistic regression produce smoother prediction surfaces compared to random forest, where predictions are not a continuous function of input variables (Brenning, 2005; Goetz et al., 2015). Several studies report better prediction performance for random forest compared with other susceptibility models (Goetz et al., 2015; Pourghasemi and Rahmati, 2018). However, in the present study, this was reversed when random forest models were fitted and tested with landslide inventories from different study areas (Table 6). Under these testing conditions, random forest underperformed logistic regression in 22 out of 24 comparisons. This suggests the random forest models are prone to over-fitting, also noted by Brenning (2005), resulting in poorer performance when fitting and testing with landslide data from different study areas.

Landslide susceptibility models fitted using manual inventory data consistently outperformed those using OBIA-derived data. This difference in performance reflects the number of false positives in the OBIA landslide inventories, particularly for the Waikato and Hawke's Bay study areas. We applied no manual refinement to semi-automated mapping results to provide a reproducible base measure of susceptibility model performance when using OBIA-derived inventories. Nonetheless, despite user's accuracies ranging 45–55% for OBIA mapping, the

Table 6
Comparison of mean AUC values based on cross-validation for logistic regression (LR) and random forest (RF) models: a) fitted with event landslide inventories from study areas 1–3 and tested with multi-temporal data (combined study areas 4–6) versus b) fitted with combined multi-temporal data from study areas 4–6 and tested with event inventories (study areas 1–3). P refers to the percentage of presence/absence points for which input variable data is available to enable prediction and testing.

Model validation	Input variables	Waikato			Whanganui			Hawke's Bay			Combined events		
		LR	RF	P	LR	RF	P	LR	RF	P	LR	RF	P
a. Event-based models tested with combined multi-temporal data	T	0.70	0.68	100	0.70	0.69	100	0.70	0.69	100	0.70	0.68	100
	T + LC	0.73	0.70	90	0.73	0.71	90	0.72	0.71	99	0.74	0.71	99
	T + LC + R	0.71	0.67	37	0.71	0.66	37	0.71	0.69	97	0.73	0.70	97
b. Multi-temporal-based model tested with event data	T	0.66	0.64	100	0.65	0.64	100	0.72	0.70	100	0.68	0.67	100
	T + LC	0.70	0.68	96	0.79	0.76	100	0.74	0.73	86	0.74	0.72	90
	T + LC + R	0.68	0.66	70	0.74	0.72	20	0.62	0.64	49	0.65	0.66	45

relative reduction of 6–11% in model predictive performance based on OBIA (AUC = 0.63–0.75) versus manual (AUC = 0.67–0.81) landslide inventories is low by comparison. This suggests landslide susceptibility analyses may be relatively insensitive to moderate classification error in semi-automated mapping, at least when using large inventories. This supports use of OBIA techniques for the acquisition of landslide inventory data. We recommend that OBIA form part of a hybrid approach combining semi-automated mapping with manual refinement (Hölbling et al., 2016). The latter should involve a systematic process for checking and refining outputs along with reporting of the resulting improvement in classification accuracy. The use of OBIA has considerable potential benefit in reducing the time and cost of acquiring landslide inventories, particularly for regional landslide events where the area affected and number of landslides are significant barriers to the acquisition of data required for susceptibility modelling.

Multi-temporal landslide inventories are generally considered the preferred source of landslide information for susceptibility modelling (Reichenbach et al., 2018). However, we did not find consistent improvement in model predictive performance when using the multi-temporal versus event inventories. This lack of distinct improvement in model performance may reflect the event-scale landslide densities, which, if sufficiently high, could reduce the relative benefit of using a multi-temporal inventory for susceptibility modelling. Moreover, we found that the smaller multi-temporal study areas were less representative of the wider range of terrain captured in the event study areas. This issue of representativeness partly reflects the greater time and cost involved in acquiring and processing repeated aerial photographs and mapping landslide scars for multiple time periods, thus constraining the number and size of multi-temporal study areas used to build landslide inventories that may limit the potential wider utility for susceptibility prediction.

In comparing models fitted and tested with landslide data from different study areas, a reduction in predictive performance was observed with the inclusion of rock type, irrespective of the model or inventory tested (Table 6). This reduction in performance illustrates model dependency on the characteristics of specific study areas that may limit potential spatial transferability without incurring a significant loss in performance (cf. Petschko et al., 2014). Rock type is effectively a proxy for potential differences in soil properties such as shear strength and hydraulic conductivity that may directly affect susceptibility to shallow landsliding (Crozier et al., 1980) but which lack spatial representation. Differences in these properties of soils formed on apparently equivalent rock types between study areas may have contributed to the reduced performance, as well as potential errors in the mapping of rock type. These differences may occur in part due to contrasts in weathering intensity that produce a range in soil properties for the same rock type (cf. Dixon et al., 2016). This finding indicates the importance of representing the same mapped rock type across multiple study areas to capture this variability in the absence of relevant data on soil properties from higher resolution mapping for use in susceptibility modelling.

We recognise other opportunities may stem from using multi-temporal inventories for investigating rainfall-initiated landslide susceptibility, such as testing long-term model performance and assessing possible changes in the distribution of landslides in a study area over time (Reichenbach et al., 2018). In contrast, event inventories present a challenge for susceptibility modelling by reflecting patterns in individual storm rainfall as well as landscape controls on the landslide response. However, selection of event study areas with knowledge of storm rainfall patterns and impacts reduces the extent to which a lack of landslide response may be incorrectly attributed to low susceptibility rather than insufficient rainfall. The present study demonstrates that susceptibility modelling with event inventories can produce comparable predictive performance to models using multi-temporal records. This finding may depend on a) event and multi-temporal inventories exhibiting comparable landslide densities and b) negligible land cover change over the interval covered by the multi-temporal inventory. Recent event inventories may also benefit from availability of VHR imagery and data for explanatory variables that are temporally consistent with the mapped storm event (e.g. repeated land cover mapping) as well as spatial information on storm rainfall magnitude-intensity (e.g. from rain radar).

5. Conclusion

We present the first comparison of landslide susceptibility model performance based on manual versus OBIA derived landslide inventories. Evaluation of OBIA for semi-automated landslide mapping showed mixed results in terms of classification accuracy, where producer's and user's accuracies ranged 62–81 and 45–55%, respectively. These results contain no manual refinement and thus provide a reproducible base measure of OBIA classification performance. Despite these moderate accuracies, the relative reduction of 6–11% in predictive performance of the susceptibility models based on OBIA versus manual inventories was low by comparison, and the spatial patterns in modelled susceptibility were generally similar. For future applications, we recommend a hybrid approach combining OBIA with a systematic manual refinement procedure to improve on this base level of performance.

The random forest model produced slightly better prediction performance compared with logistic regression when tested within the same study area. However, this was reversed and logistic regression mostly outperformed random forest when the models were fitted and tested with data from different study areas. Model predictive performance was comparable for event versus multi-temporal records based on cross-validation within the same study area. This may be explained by the density of shallow landslides within event study areas that reduced an expected improvement in model performance when using a multi-temporal inventory. In contrast, fitting models with multi-temporal data and testing with event datasets tended to outperform fitting with event data using the combination of topographic and land cover input variables, but this was reversed with inclusion of rock type. This may reflect smaller multi-temporal study areas that are less representative compared to the combined event inventories that span a wider range of terrain.

Our results highlight both the challenges associated with semi-automated landslide detection over large areas as well as the opportunity to use OBIA for efficient data collection without necessarily compromising the resulting susceptibility maps. Given the significant time and cost impediments to the preparation of multi-temporal landslide inventories, we see future targeted acquisition of multi-event landslide inventories from different locations as a viable alternative approach. This could enhance landslide susceptibility modelling by using the latest VHR satellite imagery in combination with hybrid OBIA-manual mapping to represent a wider range of terrain than would be possible with typically fewer and smaller multi-temporal datasets based on historic imagery.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the New Zealand Ministry of Business, Innovation and Employment research program “Smarter Targeting of Erosion Control (STEC)” (Contract C09X1804) and the Strategic Science Investment Fund (SSIF) allocated to Landcare Research. The Clean Water Productive Land program (contract no. C10X1006) funded data collection for the Pohangina, SE Pahiatua and Mangamaire study sites. We acknowledge Hawke's Bay Regional Council for providing satellite and aerial imagery for analysis. We thank Les Basher and two anonymous reviewers for providing helpful comments on the manuscript.

References

Amato, G., Eisank, C., Castro-Camilo, D., Lombardo, L., 2019. Accounting for covariate distributions in slope-unit-based landslide susceptibility models. A case study in the alpine environment. Eng. Geol. 260, 105237.

Basher, L.R., 2013. Erosion processes and their control in New Zealand. In Dymond JR ed. Ecosystem Services in New Zealand – Conditions and Trends. Lincoln, New Zealand: Manaaki Whenua Press, pp. 363–374.

Basher, L., Spiekermann, R., Dymond, J., Herzig, A., Hayman, E., Ausseil, A.-G., 2020. Modelling the effect of land management interventions and climate change on sediment loads in the Manawatū-Whanganui region. N. Z. J. Mar. Freshw. Res. 54, 490–511.

Betts, H., Basher, L., Dymond, J., Herzig, A., Marden, M., Phillips, C., 2017. Development of a landslide component for a sediment budget model. Environ. Model Softw. 92, 28–30.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS J. Photogramm. Remote Sens. 65, 2–16.

Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Feitosa, R.Q., van der Meer, F., van der Werff, H., van Coillie, F., Tiede, D., 2014. Geographic object-based image analysis - towards a new paradigm. ISPRS J. Photogramm. Remote Sens. 87, 180–191.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Brenning, A., 2005. Spatial prediction models for landslide hazards: review, comparison and evaluation. Nat. Hazards Earth Syst. Sci. 5, 853–862.

Burnett, N., Meyer, G.A., McFadden, L.D., 2008. Aspect-related microclimatic influences on slope forms and processes, northeastern Arizona. J. Geophys. Res. 113, F03002.

Catani, F., Lagomarsino, D., Segoni, S., Tofani, V., 2013. Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. Nat. Hazards Earth Syst. Sci. 13, 2815–2831.

Chen, W., Pourghasemi, H.R., Kornejady, A., Zhang, N., 2017. Landslide spatial modeling: introducing new ensembles of ANN, MaxEnt, and SVM machine learning techniques. Geoderma 305, 314–327.

Conoscenti, C., Rotigliano, E., Cama, M., Caraballo-Arias, N., Lombardo, L., Agnesi, V., 2016. Exploring the effect of absence selection on landslide susceptibility models: a case study in Sicily, Italy. Geomorphology 261, 222–235.

Crozier, M.J., 1996. Runout behaviour of shallow, rapid earthflows. Zeitschrift fur Geomorphologie Suppl.-Bd 105, 35–48.

Crozier, M.J., 2005. Multiple-occurrence regional landslide events in New Zealand: hazard management issues. Landslides 2, 247–256.

Crozier, M.J., 2010. Landslide geomorphology: an argument for recognition, with examples from New Zealand. Geomorphology 120, 3–15.

Crozier, M.J., 2017. A proposed cell model for multiple-occurrence regional landslide events: implications for landslide susceptibility mapping. Geomorphology 295, 480–488.

Crozier, M.J., Eyles, R.J., Marx, S.L., McConchie, J.A., Owen, R.C., 1980. Distribution of landslips in the Wairarapa hill country. N. Z. J. Geol. Geophys. 23, 5–6.

De Rose, R.C., 2013. Slope control on the frequency distribution of shallow landslides and associated soil properties, North Island, New Zealand. Earth Surface Processes & Landforms 38, 365–371.

Dixon, J.L., Chadwick, O.A., Vitousek, P.M., 2016. Climate-driven thresholds for chemical weathering in postglacial soils of New Zealand. Journal of Geophysical Research: Earth Surface 121, 1619–1634.

Dominati, E.J., Mackay, A., Lynch, B., Heath, N., Millner, I., 2014. An ecosystem services approach to the quantification of shallow mass movement erosion and the value of soil conservation practices. Ecosystem Services 9, 204–215.

Dymond, J.R., Ausseil, A.-G., Shepherd, J.D., Buettner, L., 2006. Validation of a region-wide model of landslide susceptibility in the Manawatu-Wanganui region of New Zealand. Geomorphology 74, 70–79.

Dymond, J.R., Betts, H.D., Schierlitz, C.S., 2010. An erosion model for evaluating regional land-use scenarios. Environ. Model Softw. 25, 289–298.

Dymond, J.R., Herzig, A., Betts, H., Marden, M., Phillips, C., Basher, L., 2013. Parameterisation of SedNetNZ for the Manawatu catchment. Landcare Research Contract Report no. C10X1006 prepared for AgResearch and the Ministry for Business, Innovation and Employment.

Dymond, J.R., Herzig, A., Basher, L., Betts, H.D., Marden, M., Phillips, C.J., Ausseil, A.-G., Palmer, D.J., Clark, M., Roygard, J., 2016. Development of a New Zealand SedNet model for assessment of catchment-wide soil-conservation works. Geomorphology 257, 85–93.

Gao, J., Maro, J., 2010. Topographic controls on evolution of shallow landslides in pastoral Wairarapa, New Zealand, 1979–2003. Geomorphology 114, 373–381.

Ghestem, M., Sidle, R.C., Stokes, A., 2011. The influence of plant root systems on subsurface flow: implications for slope stability. Bioscience 61, 869–879.

Glade, T., 1998. Establishing the frequency and magnitude of landslide-triggering rainstorm events in New Zealand. Environ. Geol. 35, 160–174.

Goetz, J.N., Brenning, A., Petschko, H., Leopold, P., 2015. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modelling. Comput. Geosci. 81, 1–11.

Guzzetti, F., Reichenbach, P., Ardizzone, F., Cardinali, M., Galli, M., 2006. Estimating the quality of landslide susceptibility models. Geomorphology 81, 166–184.

Guzzetti, F., Mondini, A.C., Cardinali, M., Fiorucci, F., Santangelo, M., Chang, K.-T., 2012. Landslide inventory maps: new tools for an old problem. Earth Sci. Rev. 112, 42–66.

Harrell, F.E., 2001. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer-Verlag, New York.

Heckmann, T., Gegg, K., Gegg, A., Becht, M., 2014. Sample size matters: investigating the effect of sample size on a logistic regression susceptibility model for debris flows. Nat. Hazards Earth Syst. Sci. 14, 259–278.

Hijmans, R.J., 2019. raster: Geographic Data Analysis and Modeling. R package version 2.9-23.https://CRAN.R-project.org/package=raster.

Hölbling, D., Friedl, B., Eisank, C., 2015. An object-based approach for semi-automated landslide change detection and attribution of changes to landslide classes in northern Taiwan. Earth Sci. Inf. 8, 327–335.

Hölbling, D., Betts, H., Spiekermann, R., Phillips, C., 2016. Identifying spatio-temporal landslide hotspots on North Island, New Zealand, by analysing historical and recent aerial photography. Geosciences 6, 48.

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied Logistic Regression, 3rd edn. Hoboken, NJ: Wiley Series in Probability and Statistics. John Wiley & Sons.

Inbar, A., Nyman, P., Rengers, F.K., Lane, P.N.J., Sheridan, G.J., 2018. Climate dictates magnitude of asymmetry in soil depth and hillslope gradient. Geophys. Res. Lett. 45, 6514–6522.

Jones, K.E., Levick, S.R., Page, M.J., 2011. Processing and Classifying Satellite Imagery to Assess the April 2011 Storm Induced Landsliding in Hawke's Bay. GNS Science Consultancy Report vol. 2011/265. 24pp.

Krausse, M.K., Eastwood, C., Alexander, R.R., 2001. Muddied Waters: Counting the National Economic Cost of Soil Erosion and Sedimentation in New Zealand. Manaaki Whenua Press, Lincoln.

Kritikos, T., Davies, T., 2015. Assessment of rainfall-generated shallow landslide/debris-flow susceptibility and runout using a GIS-based approach: application to western Southern Alps of New Zealand. Landslides 12, 1051–1075.

Kurtz, C., Stumpf, A., Malet, J.-P., Gançarski, P., Puissant, A., Passat, N., 2014. Hierarchical extraction of landslides from multiresolution remotely sensed optical images. ISPRS J. Photogramm. Remote Sens. 87, 122–136.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22.

Liu, J.-K., Shih, P.T.Y., 2013. Topographic correction of wind-driven rainfall for landslide analysis in central Taiwan with validation from aerial and satellite optical images. Remote Sens. 5, 2571–2589.

Lombardo, L., Mai, P.M., 2018. Presenting logistic regression-based landslide susceptibility results. Eng. Geol. 244, 14–24.

Martha, T.R., Kerle, N., Jetten, V., van Westen, C.J., Kumar, K.V., 2010. Characterising spectral, spatial and morphometric properties of landslides for semi-automated detection using object-orientated methods. Geomorphology 116, 24–36.

Naimi, B., 2015. Uncertainty Analysis for Species Distribution Models. R Software Package.

Newsome, P.F.J., Wilde, R.H., Willoughby, E.J., 2008. Land resource information system spatial data layers: data dictionary. Palmerston North, Landcare Research New Zealand Ltd. http://digitallibrary.landcareresearch.co.nz/cdm/ref/collection/p20022coll14/id/67

Petschko, H., Bell, R., Leopold, P., Heiss, G., Glade, T., 2013. Landslide inventories for reliable susceptibility maps in Lower Austria. In: Margottini, C., Canuti, P., Sassa, K. (Eds.), Landslide Science and Practice. Volume 1: Landslide Inventory and Susceptibility and Hazard Zoning. Springer, pp. 281–286.

Petschko, H., Brenning, A., Bell, R., Goetz, J., Glade, T., 2014. Assessing the quality of landslide susceptibility maps – case study Lower Austria. Nat. Hazards Earth Sys. Sci. 14, 95–118.

Petschko, H., Bell, R., Glade, T., 2015. Effectiveness of visually analyzing LiDAR DTM derivatives for earth and debris slide inventory mapping for statistical susceptibility modeling. Landslides 13, 857–872.

Phillips, C.J., Marden, M., 2005. Reforestation Schemes to Manage Regional Landslide Risk. Landslide Hazard and Risk. John Wiley & Sons Ltd, Hoboken, NJ.

Pourghasemi, H.R., Rahmati, O., 2018. Prediction of the landslide susceptibility: which algorithm, which precision? Catena 162, 177–192.

Pradhan, B., 2013. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Comput. Geosci. 51, 350–365.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna https://www.R-project/org/.

Reichenbach, P., Rossi, M., Malamud, B.D., Mihir, M., Guzzetti, F., 2018. A review of statistically-based landslide susceptibility models. Earth Sci. Rev. 180, 60–91.

Reid, L.M., Page, M.J., 2002. Magnitude and frequency of landsliding in a large New Zealand catchment. Geomorphology 49, 71–88.

Rosser, B., Dellow, S., Haubrock, S., Glassey, P., 2017. New Zealand's national landslide database. Landslides 14, 1949–1959.

Rossi, M., Guzzetti, F., Reichenbach, P., Mondini, A.C., Peruccacci, S., 2010. Optimal landslide susceptibility zonation based on multiple forecasts. Geomorphology 114, 129–142.

Ruff, M., Czurba, K., 2008. Landslide susceptibility analysis with a heuristic approach in the Eastern alps (Vorarlberg, Austria). Geomorphology 94, 314–324.

Schicker, R., Moon, V., 2012. Comparison of bivariate and multivariate statistical approaches in landslide susceptibility mapping at a regional scale. Geomorphology 161–162, 40–57.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., 2005. ROCR: visualizing classifier performance in R. Bioinformatics 21, 7881.

Stokes, A., Douglas, G., Fourcaud, T., Giadrossich, F., Gillies, C., Hubble, T., Kim, J., Loades, K., Mao, Z., McIvor, I., Mickovski, S., Mitchell, S., Osman, N., Phillips, C., Poesen, J., Polster, D., Preti, F., Raymond, P., Rey, F., Schwarz, M., Walker, L.R., 2014. Ecological mitigation of hillslope instability: ten key issues facing researchers and practitioners. Plant Soil 377, 1–23.

Trustrum, N.A., Gomez, B., Page, M.J., Reid, L.M., Hicks, D.M., 1999. Sediment production, storage and output: the relative role of large magnitude events in steepland catchments. Z. Geomorphol. N.F. Supplement 115, 71–86.

van Westen, C.J., Castellanos, E., Kuriakose, S.L., 2008. Spatial data for landslide susceptibility, hazard, and vulnerability assessment: an overview. Eng. Geol. 102, 112–131.

Youssef, A.M., Pourghasemi, H.R., Pourtaghi, Z.S., Al-Katheeri, M.M., 2016. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. Landslides 13, 839–856.