Supplementary Material

Appendix S1. LiDAR drain modelling methodology. This text expands on the subsection "Creating and validating LiDAR drain layer" in the Methods section of this manuscript.

Training points

Training points for the random forest model were created in ArcMAP 10.5 using the digital elevation model and base layer New Zealand aerial imagery. We digitised 166 locations and classified the landform at these locations into 18 classes of landscape position in and adjacent to drains, including four drain specific classes. These locations were not ground-truthed; we suggest more locations, and more ground-truthed locations, would provide a more accurate (and precise) outcome.

We ran a classification model using the randomForest R package (Liaw & Wiener 2002) and created an initial raster model. Data were split into training data with 124 sites and testing data with 41 sites to help assess the model fit for the evaluation. We chose eight layers to use in the random forest model to predict whether a pixel was a drain or not. These layers related to landform relief and are tabulated in Appendix S2. We were aware of the risk of overfitting and autocorrelation in using all these layers but considered it appropriate for the task at hand as we were not seeking to predict outside of the study area.

Covariate layers were more numerous than those described in Roelens et al (2018); we found this necessary due to the greater diversity of near-drain topography compared with other studies, which led to the inclusion of an excessive number of false-positives (where drains were indicated where no drain was apparent).

We used the ArcGIS model builder to create what is known as a vectorisation and segmentation workflow, which converted the pixel-based random forest raster output to polygons of identified drains. These were visually checked against imagery and landform covariates. The final LiDAR-based catchment-scale drains layer was created by clipping to a mask that represented the area of interest and excluded areas at the edge of the LiDAR layer, where the LiDAR data were not adequate to create a drains layer. As a final clean up step, areas of drains that overlapped with the NZ lakes dataset (https://data.linz.govt.nz/layer/50293-nz-lake-polygons-topo-150k/) were erased, and small polygons were eliminated. To eliminate small polygons we firstly buffered the drains by 10 m and then assigned the newly unioned polygons to a 'cluster'; we selected clusters (groups of nearby polygons) that were > 300 m^2; polygons that were attributed to these clusters were retained.

We then validated the model by selecting 120 random points within the Waituna Lagoon catchment study area and creating circles of 5 ha with the point in the centre (Appendix S3). We include only those that were fully inside the study area (n = 114); this meant we covered 570 ha in the validation exercise. Within each circle we manually digitised the centreline of all the drains we could detect using the ESRI basemap NZ Imagery layer and two of the covariate data layers from the Drains Random Forest model: ZRel and Topographic Protection Index (TPI). ZRel is a measure of the relative height of a pixel and the mean of its surrounding pixels in a small (5 × 5 pixel circle) neighbourhood. TPI is essentially the same calculation as ZRel but run over a 100 m radius instead of 15 m (in these instances). For each digitised drain centreline we measured and recorded the half-width of the drain from edge to edge of the apparent cut; and used this distance to buffer the centre line to create a polygon. This gave us the same kind of data (drains represented by polygons) as provided by the model. As noted in the main text, we assess the model results using the following metrics:

(1) Overall agreement: the area of drain mapped as drain by both methods, added to the area mapped as 'not drain' by both datasets, divided by the total area considered (effectively 114 × 5 ha).

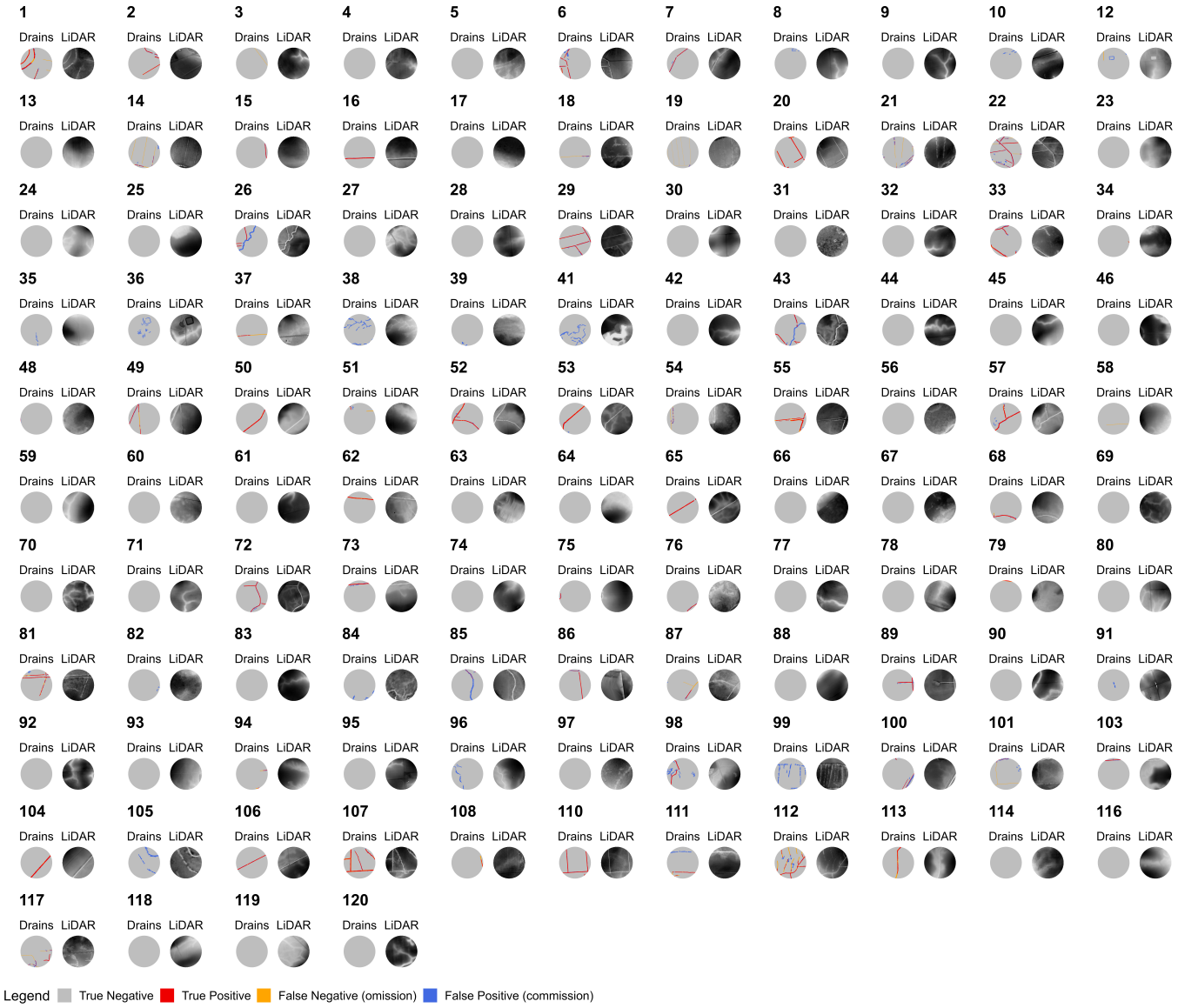
(2) Overall omission: the amount of drain not detected by the model. Can be expressed as area (total area omitted/total area of false negative), or as a percent (area omitted, out of the validated area of drain). This formulation describes the percentage of drain missed.

(3) Overall commission: the amount of drain identified by the model, that was not identified by the validation method. Can be expressed as area (total area of commission/total area of false positive), or as a percent. In this case, the percentage is calculated using the area of commission divided by the total area identified as drain by the model (including the false positive). This effectively describes how much of the area the model incorrectly identified as drain. This formulation avoids dividing by zero (model identifies some drain in a polygon, but validation exercise does not).

Appendix S2. Covariate layers used in the preliminary LiDAR evaluation of the case study catchment.

Layer name	Explanation
SL2109_Zrel	Local Relative Elevation
TopographicProtectionIndex_25cm	Topographic Protection Index
SL2109_A02_Waituna_Slope_pt25mSlope_Waituna	Slope
SL2108_Covariate_MidSlopePositon	Mid Slope Position
SL2108_Covariate_NormalizedHeight	Normalized Height
SL2108_Covariate_SlopeHeight	Slope Height
SL2108_Covariate_StandardizedHeight	Standardized Height
SL2108_Covariate_ValleyDepth	Valley Depth

Appendix S3. Validation circles (numbered in bold above each panel) adjacent to the same area shown with LiDAR. Whether a model result was defined as true/false positive/negative was defined by level of agreement with a human operator digitising any visible drains as 'truth' – described further above.



References

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

Roelens J, Rosier I, Dondeyne S, Van Orshoven J, Diels J 2018. Extracting drainage networks and their connectivity using LiDAR data. Hydrological Processes 32(8): 1026–1037.